

Assembly and Repeat Annotation of the *Nothobranchius furzeri* Genome

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät
der Friedrich-Schiller-Universität Jena

von Diplom-Bioinformatiker Philipp Koch

geboren am 10.07.1983 in Gotha

Die vorliegende Arbeit wurde in der Zeit von Januar 2011 bis Februar 2016 am Leibniz-Institut für Alternsforschung - Fritz-Lipmann-Institut (FLI) in Jena angefertigt.

Gutachter:

1. Prof. Dr. Christoph Englert, Forschungsgruppe Molekulare Genetik, Leibniz-Institut für Alternsforschung - Fritz-Lipmann-Institut, Jena
2. Dr. Reinhard Guthke, Forschungsgruppe Systembiologie, Leibniz-Institut für Naturstoff-Forschung und Infektionsbiologie - Hans-Knöll-Institut, Jena
3. Prof. Dr. Thomas Wiehe, Institut für Genetik, Universität zu Köln, Köln

Datum der Verteidigung: 19.05.2016

Table of Contents

Table of Contents	I
Abbreviations	IV
List of Figures	V
List of Tables	VI
Zusammenfassung	VII
Abstract	IX
1 Introduction	1
1.1 Genome Sequencing and Assembly	1
1.1.1 Genome Composition	1
1.1.2 Sequencing Technologies	2
1.1.2.1 First Generation Sequencing	2
1.1.2.2 Second Generation Sequencing	3
1.1.2.3 Third Generation Sequencing	4
1.1.3 <i>De novo</i> Genome Assembly Methods	5
1.1.3.1 Overlap Layout Consensus	6
1.1.3.2 De Bruijn Graph	6
1.1.4 Scaffolding	7
1.1.5 Optical Mapping	7
1.1.6 Genetic Mapping	8
1.1.7 Synteny Scaffolding	9
1.2 Repetitive Elements in Genomes	9
1.2.1 Tandem Repeats	9
1.2.2 Transposable Elements	10
1.2.2.1 Class I - Retrotransposons	10
1.2.2.2 Class II - DNA Transposons	11
1.2.3 Gene Families and Segmental Duplications	11
1.2.4 Identification of Repetitive Elements	12
1.2.4.1 Experimental Repeat Identification	12
1.2.4.2 Similarity-Based Repeat Identification	13
1.2.4.3 <i>De novo</i> Repeat Identification	13
1.2.4.4 Signature-Based Repeat Identification	14
1.2.4.5 Integrated Repeat Identification and Annotation Pipelines	14
1.2.4.6 Repeat Classification Programs	14
1.3 <i>Nothobranchius furzeri</i>	15
1.3.1 <i>N. furzeri</i> is a Model Organism in Aging Research	15
1.3.2 Genetic and Genomic Resources	16
1.3.2.1 Initial Characterization of the <i>N. furzeri</i> Genome	16

1.3.2.2	Linkage Maps of the <i>N. furzeri</i> Genome	17
1.3.2.3	Sequencing Resources	17
1.4	Thesis Objectives	18
2	Material and Methods	19
2.1	Assembly of the <i>N. furzeri</i> Genome	19
2.1.1	DNA Sequencing	19
2.1.1.1	Sanger Sequencing	19
2.1.1.2	454 Sequencing	19
2.1.1.3	Illumina Sequencing	20
2.1.1.4	PacBio Sequencing	20
2.1.2	DNA Preparation for Optical Mapping	22
2.1.3	Data Preparation for ALLPATHS-LG	22
2.1.4	Basic Assembly with ALLPATHS-LG	22
2.1.5	Further Scaffolding with KILAPE and Additional Gap Filling	23
2.1.6	Ordering Scaffolds with Optical Mapping	23
2.1.7	Linkage Map-Based Scaffolding	24
2.1.8	Building Synteny Groups	25
2.1.9	Additional PacBio-Based Analyses	25
2.2	The Repeat Identification Pipeline RepARK	26
2.2.1	The <i>D. melanogaster</i> Genome	26
2.2.2	Available Repeat Libraries for <i>D. melanogaster</i>	26
2.2.3	Sequencing Data	27
2.2.4	Building the RepARK Repeat Libraries	27
2.2.5	Building Additional Repeat Libraries for Comparison	28
2.2.6	Completeness Calculation	28
2.2.7	Repeat Classification with TEclass	28
2.2.8	Mapping and Repeat Masking	29
2.2.9	Retrieving Known SDs and Comparison to the <i>de novo</i> Repeat Consensi	29
2.3	Identification and Classification of Genomic <i>N. furzeri</i> Repeats	29
2.3.1	Genome Assembly-Based Library Creation	29
2.3.2	Read Data-Based Library Creation	29
2.3.3	Size Selection and Redundancy Reduction of Repeat Libraries	30
2.3.4	Repeat Classification	30
2.3.5	Genome Assembly Repeat Annotation	30
2.3.6	Repeat Analysis of PacBio-Filled Gaps and Completeness Estimation of the Genome Assembly	31
3	Results	33
3.1	Assembling a Reference Sequence of the <i>N. furzeri</i> Genome	33
3.1.1	Sequencing and Data Preparation	33
3.1.2	Genome Assembly	33
3.1.2.1	Basic Assembly with ALLPATHS-LG - Assembly A	34
3.1.2.2	Further Scaffolding with KILAPE and Gap Filling - Assembly B	34
3.1.2.3	Superscaffolding with Optical Maps - Assembly C	35
3.1.2.4	Linkage Map-Based Scaffolding - Assembly D	37
3.1.2.5	Synteny-Based Scaffolding - Assembly E	39
3.1.2.6	Summary and Availability of the Genome Assembly	41

3.2	Developing RepARK – A New Method for <i>de novo</i> Repeat Analysis	42
3.2.1	The RepARK Pipeline	42
3.2.2	Evaluation of RepARK Based on the <i>D. melanogaster</i> Genome	43
3.2.2.1	<i>D. melanogaster</i> Repeat Library Construction	43
3.2.2.2	Basic Repeat Library Characteristics.....	44
3.2.2.3	Sensitivity	45
3.2.2.4	Identification of Putative Novel Repeats	48
3.2.2.5	Classification of <i>D. melanogaster</i> Repeats	49
3.2.3	RepARK on the Complex Human Genome	50
3.2.3.1	Human Repeat Library Construction.....	50
3.2.3.2	Detection of the Epstein-Barr Virus Genome in Human Data.....	51
3.3	Comprehensive <i>N. furzeri</i> Repeat Analysis and Genome Annotation	52
3.3.1	Basic Repeat Analysis of the Genome Assembly	52
3.3.2	Building a Library of <i>N. furzeri</i> Repeat Consensi	52
3.3.3	Classifying Consensi of the Combined Repeat Library	54
3.3.4	Repeat Annotation of the Genome Assembly	54
3.3.4.1	Performance of Single Repeat Libraries.....	54
3.3.4.2	Repeat Composition of the Genome Assembly	55
3.3.4.3	Evolutionary History of Transposable Elements	55
3.3.5	Determination of the Repetitive Fraction of the <i>N. furzeri</i> Genome	58
3.3.6	Repeat-Based Estimation of the Completeness of the <i>N. furzeri</i> Genome Assembly	59
4	Discussion.....	61
4.1	Strategies for the <i>de novo</i> Genome Assembly of <i>N. furzeri</i>	61
4.2	Repeat Identification and the Development of RepARK.....	68
4.3	Repeat Content of the <i>N. furzeri</i> Genome	71
5	Conclusions and Outlook.....	77
6	References	78
	Supplemental Figures	i
	Supplemental Tables	v
	Acknowledgements	vii
	Selbstständigkeitserklärung.....	viii

Abbreviations

A

A adenine
ATP adenosine triphosphate

B

BAC bacterial artificial chromosome
bp base pair

C

C Celsius, cytosine
CDS coding sequence
cM centi Morgan
CPU central processing unit

D

DBG de Bruijn graph
ddNTP 2',3'-dideoxynucleotide
DIRS Dictyostelium intermediate repeat sequence
DNA deoxyribonucleic acid
dNTP 2'-deoxynucleotide
DR dispersed repeat

E

EBV Epstein-Barr virus
et al. et alii (and others)

F

FISH fluorescence *in situ* hybridization
FLI Leibniz-Institute for Aging - Fritz-Lipmann-Institute

G

G guanine
Gb giga bases
GB gigabyte
GRZ Gonarezhou
GSC genetic scaffold

H

h hour

I

ID identifier

L

LG linkage group
LINE long interspersed nuclear elements
LTR long terminal repeat

M

Mb mega bases
MITE miniature inverted repeat
MZM Mozambique

N

N any base (A, C, G, T, U)
NGS next generation sequencing
nt nucleotide
Nxx the length of a sequence above which xx% of the total assembly size is represented

O

OLC overlap layout consensus

P

PCR polymerase chain reaction
PLE penelope-like elements

R

RAD restriction site associated DNA
RNA ribonucleic acid
ROI read of insert
rRNA ribosomal RNA
RT reverse transcriptase

S

SD standard deviation, segmental duplication
SGR synteny group
SINE short interspersed nuclear elements
SMRM single molecule restriction map
SMRT Single Molecule Real-Time
SNP single nucleotide polymorphism
SNV single nucleotide variation
SRA Short Read Archive
SSR simple sequence repeat
STR short tandem repeat

T

T thymine
TE transposable element
TIR terminal inverted repeat
TR tandem repeat
TRF Tandem Repeats Finder
tRNA terminal inverted repeat
TSD target site duplication

U

U uracil

V

vs. versus

W

WGS whole genome shotgun

List of Figures

Figure 1: Alignment of Optical Mapping and Sequencing Data for Part of <i>N. furzeri</i> superscaffold00010.	36
Figure 2: Annotation of Maptig Ends by OpGen.	37
Figure 3: Construction of Genetic Scaffolds.	38
Figure 4: Assembly Overview Including 19 Synteny Groups.	40
Figure 5: Cumulative Length of the Assemblies A-E.	42
Figure 6: Outline of the RepARK Pipeline.	43
Figure 7: 31-Mer Coverage Histograms for Simulated and Real <i>D. melanogaster</i> Illumina Reads.	44
Figure 8: Total Length of the <i>D. melanogaster</i> Repeat Libraries.	46
Figure 9: Repetitive Genomic Fraction of the <i>D. melanogaster</i> Reference Genome.	46
Figure 10: Completeness of 212 <i>D. melanogaster</i> RepBase Repeats Separated by their Main TE Class.	47
Figure 11: Completeness of the 249 <i>D. melanogaster</i> RepBase Repeats in the Four RepARK Libraries.	48
Figure 12: Genomic Fraction of Known SDs and their Representation in the Repeat Libraries.	49
Figure 13: Genomic Fraction of TEclass-Classified Repeat Consensi.	50
Figure 14: High Confidence Alignments of Human RepARK Consensi to the EBV Genome.	51
Figure 15: Distribution of the Occurrence of 31-mers that Map to the EBV Genome.	52
Figure 16: Repeat Classification of the CombinedLib.	54
Figure 17: Evolutionary History of the Major TE Classes in the <i>N. furzeri</i> Genome Assembly.	56
Figure 18: Evolutionary History of TE superfamilies in the <i>N. furzeri</i> Genome Assembly.	57
Figure 19: Repeat Content in Assembled and Non-Assembled <i>N. furzeri</i> Data.	58

List of Tables

Table 1: <i>N. furzeri</i> GRZ Individuals in this Study and Their Use for Various Analyses.	19
Table 2: Sequence Data Used for Scaffolding with KILAPE, Gap Filling and Repeat Analyses.....	21
Table 3: Sequence Data Used for the ALLPATHS-LG Initial Assembly.....	22
Table 4: Available <i>D. melanogaster</i> Repeat Libraries.	27
Table 5: Sequence Data Used for the RepARK Development and Evaluation.	27
Table 6: Sequence Data Used for the ALLPATHS-LG Initial Assembly after Down-Sampling.	33
Table 7: Assembly Statistics.	34
Table 8: OpGen Scaffolds of Assembly B, Broken at Gaps >5 kb and Used for Optical Mapping.....	35
Table 9: Maptigs Obtained by SMRM <i>de novo</i> Assembly.....	35
Table 10: Construction Scheme of the SGRs.	41
Table 11: <i>D. melanogaster</i> Repeat Library Metrics from Simulated NGS Reads.	45
Table 12: <i>D. melanogaster</i> Repeat Library Metrics from Real Data.	45
Table 13: BLAT Mappings of Repeat Libraries to the <i>D. melanogaster</i> Genome.....	46
Table 14: Human Repeat Library Metrics and Mapping Results Against the Human Reference Sequence.....	50
Table 15: Repeat Libraries of <i>N. furzeri</i>	53
Table 16: Repeat Fraction of the Genome Assembly Annotated by Different Repeat Libraries of <i>N. furzeri</i>	55
Table 17: Repeat Composition of the <i>N. furzeri</i> Genome Assembly	55
Table 18: Filled Gaps Using PacBio Sequences.	60
Table 19: Gaps of the Genome Assembly.....	60

Zusammenfassung

In der Altersforschung an Vertebraten sind die relativ langen Lebensspannen der derzeitigen Modellorganismen ein Hindernis bei der Durchführung von Experimenten. Der Türkise Prachtgrundkärpfling *Nothobranchius furzeri* lebt in saisonalen Teichen im Südosten Afrikas und weist eine sehr kurze Lebensspanne auf (4-12 Monate). Die Tiere dieser Spezies zeigen ein schnelles Wachstum, eine frühe Geschlechtsreife und typische Biomarker des Alterns. Fische aus Gebieten mit verschiedenen langen Regenzeiten unterscheiden sich auch in ihrer Lebensspanne. Im Labor bleiben diese Unterschiede erhalten und lassen darauf schließen, dass das Merkmal Lebensspanne im Erbgut verankert sein muss. Diese Eigenschaften machen *N. furzeri* zu einem geeigneten Modellorganismus für die Altersforschung.

Eine der wichtigsten Ressourcen zur Untersuchung der Biologie und Genetik des Alterns von *N. furzeri* ist die Kenntnis seiner Genomsequenz. In der vorliegenden Arbeit beschreibe ich den mehrstufigen Prozess zur Schaffung dieser Ressource, zu dem ich maßgeblich beitrug. Dazu wurden zunächst Sequenzdaten, die mit Geräten der zweiten Sequenziergeneration erzeugt wurden, mit externen und selbstentwickelten Programmen assembliert. Anschließend wurden weitere Ressourcen wie optische und genetische Karten sowie Syntänievergleiche herangezogen um eine qualitativ hochwertige Genomassemblierung zu erzielen, die eine Größe von 1.24 Gb (N50 57.4 Mb) aufweist und entsprechend der 19 *N. furzeri* Chromosomen gegliedert ist.

Vorangegangene Arbeiten hatten gezeigt, dass das *N. furzeri* Genom einen hohen Anteil und eine außergewöhnliche Zusammensetzung repetitiver Sequenzen aufweist. Da diese die Genomassemblierung erschweren, habe ich sie gesondert und detailliert untersucht. Zu diesem Zweck wurde die Softwarepipeline RepARK entwickelt, die repetitive Elemente (Repeats), basierend allein auf den Rohdaten der Hochdurchsatz-Sequenzierverfahren, erkennt. Diese Methode steht dem bisher etablierten Herangehen gegenüber, das zur Repeaterkennung von einer bereits assemblierten Genomsequenz ausgeht, in der aber Repeats zu Fehlassemblierungen führen können und unterrepräsentiert sind. RepARK unterliegt nicht diesen Verzerrungen und erstellt anhand von k-mer Häufigkeiten der Rohdaten eine Bibliothek von speziesspezifischen Konsensussequenzen der gefunden Repeats.

Indem ich RepARK und weitere Methoden angewendet habe, erzeugte ich eine umfassende Repeatbibliothek von *N. furzeri* und nutzte sie um das assemblierte *N. furzeri* Genom zu annotieren. Von den 1.24 Gb wurden 441 Mb (35.5%) als Repeats eingestuft, welche ich auf ihren Typ, ihre Häufigkeit und ihre Dynamik im Verlauf der Genomevolution untersucht habe. Die größten Anteile haben LINE Retrotransposons und DNA Transposons, wobei einige Unterklassen beider Typen Anzeichen aufweisen, sich noch immer aktiv im *N. furzeri* Genom zu vermehren.

Basierend auf dieser Repeatannotation und Daten der Einzelmolekül-Sequenzierung konnte ich abschätzen, dass die aktuelle Genomassemblierung zwar circa 90% der unikal Sequenzen enthält aber rund 60% der Repeats fehlen. Das deckt sich mit meiner Bestimmung des Repeatgehalts in unassemblierten Rohdaten, die auf einen repetitiven Anteil am *N. furzeri* Genom von 56-70% hindeuten. Diesen gilt es in Zukunft insbesondere mit Sequenzierungstechnologien der dritten Generation möglichst vollständig zu erfassen, um auf dieser Basis die Rolle von repetitiven Elementen in biologischen Prozessen wie dem Altern erforschen zu können.

Abstract

In aging research, the long lifespan of the current vertebrate model organisms challenges the feasibility of research efforts. The turquoise killifish *Nothobranchius furzeri* lives in seasonal ponds in South-East Africa and has the shortest lifespan for vertebrates in captivity known so far (4-12 months). The fish undergoes an accelerated development with a fast growth rate, shows early sexual maturity and expresses aging-related biomarkers. Fish from regions with different durations of rainy seasons also differ in their lifespan. These differences are also observed in the laboratory and suggested a genetic determination. For these reasons, *N. furzeri* has been established as a valuable new vertebrate model in aging research.

An essential pre-requisite to study the biology and genetics of *N. furzeri* aging is the knowledge of its genome sequence. In this thesis, I describe the multi-step process of building this reference sequence to which I made major contributions. Second generation sequencing data were assembled both with external and newly developed in-house programs. Using these data, subsequently optical and genetic mapping as well as synteny analyses was used to build a high-quality genome assembly of 1.24 Gb (N50 57.4 Mb) and sequences on a chromosomal scale.

Previous studies showed that the *N. furzeri* genome harbors a high fraction of repetitive sequences (repeats) with a remarkable composition. Because repeats represent a major challenge for the genome assembly, I analyzed them separately and in detail. To this end, the software pipeline RepARK was developed for the detection of repeats in high-throughput sequencing data. The approach of the program is different to established methods that require an assembled genome, in which repeats possibly lead to assembly errors or are underrepresented. RepARK is not affected by such biases and builds, based on k-mer frequencies in the dataset, a repeat library containing representative species-specific repeat consensus sequences.

Applying RepARK and additional methods, I built a comprehensive repeat library that was subsequently used to annotate the genome assembly. Of the 1.24 Gb, 441 Mb (35.5%) are annotated as repeats which I further characterized in terms of type, occurrence and their dynamics during genome evolution. I found that LINE retrotransposons and DNA transposons are most abundant in the genome assembly and subclasses of both are probably still actively proliferating in the genome.

Based on the repeat annotation of the genome assembly and data from single-molecule sequencing, I estimated that 90% of the unique sequence is contained in the assembly but about 60% of the repeats are absent. This is consistent with my repeat analyses in not-assembled data that suggest a repeat content in the *N. furzeri* genome ranging from 56% to 70%. This missing fraction needs to be resolved by further third generation sequencing efforts to study the role of repeats in biological processes such as aging.

“The field of NGS¹ development and applications is a fast-moving area of research, which makes this an exciting time for genomic studies.”
(Metzker 2010)

1 Introduction

1.1 Genome Sequencing and Assembly

Knowing the genomic sequence of an organism plays a key role for understanding determinants of its biology. Having a high-quality genome sequence available allows genome-wide experimental and computational analyses. Although the sequencing technologies improved remarkably over the last 35 years, it is still challenging to piece sequences together (“assemble”) to obtain long and error free contiguous sequences that represent chromosomes. In this section, an introduction in sequencing technologies as well as computational approaches for genome assembly and genome-wide analyses is given.

1.1.1 Genome Composition

The heritable information of organisms is contained in the nucleic acid macromolecules deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). The four main nucleobases in DNA are adenine (A), cytosine (C), guanine (G) and thymine (T). A, C and G are also present in RNA while thymine is usually substituted by uracil (U). An organism can have one or several DNA or RNA molecules which together form its genome. Even particles not classically defined as living beings, like viruses that need a cellular host for replication and survival, carry DNA or RNA. The size of genomes varies greatly between the different kingdoms in the tree of life. Viruses and prokaryotes can exist with some tens of thousands of nucleotides (nt) in their genomes whereas simple eukaryotes have millions of them and more complex eukaryotic organisms like various plants and vertebrates have genomes comprising billions of nucleotides. Surprisingly, the genome size is not directly correlated with the complexity of an organism (Mirsky and Ris 1951), which is called the C-value paradox (Thomas 1971). Because the number of genes is relatively similar in different organisms, other regions of a genome must be responsible for the genome size difference (Petrov 2001). These, so-called intergenic regions harbor sequences that occur multiple times in the genome in contrast to the unique sequence of genes and are therefore named “repetitive elements” or “repeats”.

The DNA double helix is usually an extremely large polymer structure and therefore bound to proteins that stabilize and protect it. In eukaryotes, these are histones and the respective nucleoprotein

¹ NGS: next generation sequencing, refers to high-throughput sequencing technologies.

complex is called chromatin. A chromatin fiber is coiled in different packaging levels finally constituting a chromosome. The latter are composed of euchromatic and heterochromatic regions. Euchromatin consists of genomic regions mostly (but not always) transcriptionally active and can be condensed or uncondensed, depending on the stage of the cell cycle. In contrast, heterochromatin is usually condensed, often highly repetitive in its sequence and is part of special chromosomal regions like centromeres or telomeres.

1.1.2 Sequencing Technologies

In the context of this work, sequencing is defined as the process of reading the nucleotide sequence of DNA molecules resulting in “sequencing readouts” or “reads”. Current sequencing technologies are usually categorized by “generations”. The fundamental inventions were made in the late 1970s and are commonly referred to as “first generation sequencing technology”. For more than 25 years this was the state-of-the-art technology extensively used in many sequencing efforts including the most comprehensive projects leading to the first assemblies of the human genome sequence (Lander et al. 2001; Venter et al. 2001). Due to improvements in biotechnology, microscopy and computer technology, a second generation of sequencing technologies emerged that produced many more sequences in parallel at lower costs. In the last years, a third generation arose that aims to sequence single molecules and to produce reads that are at least several thousand base pairs (bp) in length. All current technologies have in common that the analyzed DNA molecules are fragmented prior to sequencing. One can sequence either the entire DNA fraction of a biological sample (“whole genome shotgun” (WGS) sequencing) or parts of it after enrichment. Such parts can include, for example, only exons (by whole exome sequencing) or regions with protein interactions (by chromatin immuno precipitation DNA-sequencing). DNA molecules can be read from one end (single-end reads) or from both ends (paired-end and mate-pair reads). The most prominent technologies will be described in the following.

1.1.2.1 First Generation Sequencing

Although initial technologies for determining the nucleotide sequence of DNA molecules were labor intensive and yielded only few data, they opened up new perspectives in molecular biology. Here, single DNA molecules are initially amplified and separately sequenced, and later on this was done in automated processes.

Maxam-Gilbert Sequencing

In 1977, Maxam and Gilbert developed one of the first sequencing methods. After fragmentation of the DNA by restriction enzymes, each fragment is labeled at one end with a radioactive isotope of phosphorus (^{32}P). Different base-specific cleavage reactions are applied to break the molecules at either purines (the ‘A+G’ reaction), preferable at adenine (‘A>G’), at pyrimidines (‘C+T’), and at cytosines only (‘C’). These steps produce a mixture of fragments that were separated by gel electrophoresis allowing the read-out of the nucleotide sequence (Maxam and Gilbert 1977). Low

speed, usage of hazardous chemicals and modest read lengths hindered an application of the method in larger scale for genome sequencing.

Sanger Sequencing

In parallel to Maxam and Gilbert, Frederick Sanger and colleagues developed a sequencing method based on the synthesis of a complementary DNA strand using DNA polymerase, natural 2'-deoxynucleotides (dNTPs) and artificial 2',3'-dideoxynucleotides (ddNTPs). The DNA polymerase builds the new complementary strand by incorporating either dNTPs, by which the new strand is elongated, or ddNTPs, by which the polymerase reaction is terminated due to the absence of the 3'-hydroxyl group. Thereby DNA fragments of differing lengths are produced which are radioactively labeled and separated by gel electrophoresis (Sanger et al. 1977). Technology improvements led to the development of automated Sanger sequencing, where primers or ddNTPs were fluorescently labeled and automatically detected, yielding read-lengths of up to 800-1,000 bp (Metzker 2005; Hutchison 2007).

Pyrosequencing

In contrast to the sequencing methods mentioned before which either break DNA fragments chemically at distinct nucleotides or use modified dNTPs to terminate DNA synthesis and then determine fragment lengths based on radioactivity or fluorescence, pyrosequencing detects the emission of bioluminescence. Similar to the Sanger method, sequencing starts with a single DNA strand of which the complementary strand is synthesized during repeated cycles where one of the four possible dNTPs is presented to the DNA template. When the DNA polymerase incorporates one or more dNTPs into the new strand, an inorganic pyrophosphate per incorporated nucleotide is released. The produced pyrophosphate is converted to adenosine triphosphate (ATP) by ATP sulfurylase, which is then sensed by luciferase. The amount of light produced in the latter reaction is measured thus allowing the detection of the incorporated nucleotide (Ronaghi et al. 1996; Ronaghi et al. 1998).

1.1.2.2 Second Generation Sequencing

First generation sequencing technologies were constantly improved but their limitations in throughput required new approaches. In the second generation, the sequencing reactions are carried out in a highly parallel manner allowing to sequence large genomes with high coverage.

454/Roche

The 454/Roche technology (in the following referred to as "454") was the first technology of the 2nd generation and is the high-throughput version of pyrosequencing (Margulies et al. 2005). First, DNA fragments are ligated to adapters carrying universal PCR (polymerase chain reaction) primer sequences and each fragment is attached to a separate magnetic bead. Next, the DNA template on a single bead is amplified separately in an emulsion PCR. Each water droplet filled with a bead forms an isolated reaction chamber for the template amplification. Up to one million of these beads are then deposited into individual "PicoTiterPlate" wells where the pyrosequencing reactions are carried out in

parallel (Metzker 2010). The 454 GS FLX Titanium platform yielded 450 mega bases (Mb) per run with an average read length of 330 bp (Metzker 2010) which was later improved to achieve mode read lengths of 700 bp and maximum read lengths up to 1,000 bp². A major drawback of the technology is its low performance in correctly resolving homopolymer repeats (Hutchison 2007).

Solexa/Illumina

For the Solexa/Illumina (in the following referred to as “Illumina”) sequencing, the DNA fragments are attached to an adapter sequence and amplified on a glass slide (“flow-cell”) which is organized in up to 8 lanes. On the surface of a flow-cell, oligonucleotides are attached at high-density to which the free end of the individual DNA fragments can hybridize. Each fragment is amplified by a bridge PCR resulting in a dense cluster of identical DNA fragments. To initiate the sequencing reaction, a sequencing primer is hybridized to the free end adapter of the amplified fragments, which in turn allows the DNA polymerase to add nucleotides. Then, all four fluorescently labeled nucleotides with blocked 3’ ends are applied to the reaction at the same time and exactly one nucleotide per molecule gets incorporated into the newly synthesized strand. Laser light stimulates the emission of fluorescence of the incorporated nucleotides and a camera records the light emission of all clusters. To finish the cycle, a cleavage step follows, which removes the 3’ block of the nucleotide, so that a new nucleotide can be incorporated in the next cycle. The order of the different colors of a cluster determines the sequence of the read. Since its introduction, the technology was constantly improved allowing currently read length of up to 300 nt (MiSeq³) and a data output of 600 giga bases (Gb) per day (HiSeqX Ten⁴)

1.1.2.3 Third Generation Sequencing

Central to 3rd generation sequencing technologies is that single molecules are sequenced thus avoiding the amplification step of the templates. Sequencing is performed in real time, meaning that data read-out is coupled with the reaction rate and not determined by the sequencing cycles of the device.

Pacific Biosciences

The first and most prominent 3rd generation technology was commercialized by Pacific Biosciences (in the following referred to as “PacBio”) as “Single Molecule Real-Time (SMRT) Sequencing” and detects the dNTP incorporation during DNA synthesis. The sequencing reaction is carried out in so-called SMRT cells, where in nano meter-sized holes a DNA polymerase molecule is fixed. The DNA molecule is circularized at either end with special adapters to which sequencing primer bind. One such DNA template is processed by one polymerase. With every fluorescently tagged dNTP incorporated into the newly-synthesized DNA strand, the emitted light is detected by an imaging system. Each reaction is constantly observed and recorded, and the resulting sequence of light intensities over time is referred to as a polymerase read (Eid et al. 2009). Because the DNA template is circularized, the

² <http://454.com/products/gs-flx-system/index.asp>

³ <http://www.illumina.com/systems/miseq.html>

⁴ <http://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote-hiseq-x.pdf>

polymerase can read both strands repeatedly and thus passes through the same position several times during a sequencing run. The readings of the template are concatenated in a single polymerase read from which so-called subreads are extracted where each subread represents one pass through the template. These subreads can reach lengths of >20-30 kb and one SMRT cell yields currently 350 Mb of data on a PacBio RS II system (Kim et al. 2014). Another system was recently presented by PacBio (Sequel⁵) with a 7-fold increase in yield per SMRT cell. Currently, a major disadvantage of SMRT sequencing is the high per-base error rate of approximately 15% in the raw polymerase reads which challenges sequence alignments or *de novo* sequence assemblies (Chin et al. 2013; Lee et al. 2014). To address this issue, methods were developed that either use 2nd generation reads for error correction (Koren et al. 2012; Ribeiro et al. 2012) or pre-assemble overlapping subreads to create accurate seeds which are subsequently extended to longer sequences (Chin et al. 2013).

Oxford Nanopore

This technology retrieves sequence information upon traversing a DNA molecule through a protein pore of a few nanometers in diameter. The pore is incorporated into a membrane and an electric potential is established across the membrane. An ionic current flowing only through the pore is constantly monitored (Scheicher et al. 2012). Molecules that pass the pore cause characteristic disruptions in this current. For DNA sequencing, libraries of long molecules are prepared and the double-stranded molecules subsequently bound to a processing enzyme that approaches the nanopore, unwinds the DNA, and guides a single strand through the pore. As the nucleotides pass through the pore, a characteristic profile of current is recorded from which the DNA sequence is deduced. Multiple nanopores are arranged on array chips to sequence multiple molecules at a time (Clarke et al. 2009). The first instrument released by Oxford Nanopore Technologies (MinION⁶) is a compact device with one array chip applicable for small samples but large-scale solutions are also announced by the company.

The MinION device reaches a medium and maximum read length of around 5 kb and 66 kb, respectively (Ashton et al. 2014). In a first study, the identity of mapped reads to a respective reference was reported to be below 10% (Mikheyev and Tin 2014). Improvements of technology and chemicals resulted in accuracies of up to 80% with a maximum sequence yield of 400 Mb per run (Ashton et al. 2014; Quick et al. 2014). Although the rate of errors is still high and the yield is relatively low, MinION reads were successfully combined with Illumina reads in a hybrid approach to assemble the genome of a *Salmonella* strain (Ashton et al. 2014).

1.1.3 *De novo* Genome Assembly Methods

The reconstruction of the genomic sequence of an organism is referred to as genome assembly. Like in a jigsaw puzzle where a picture is reconstructed from small pieces, genome sequences are also

⁵ <http://www.pacb.com/products-and-services/pacbio-systems/sequel/>

⁶ <https://www.nanoporetech.com/products-services/minion-mki>

assembled from relatively short sequence fragments. This fragmentation is required because no technology is currently available (or even announced) that allows to read-out an entire chromosome from one end to the other. In contrast to a reference-based genome assembly, which benefits from an available assembled genomic reference sequence that is used as a backbone, a *de novo* assembly connects overlapping reads to longer contiguous sequences, so-called “contigs” that represent a consensus of the respective reads. The repetitive nature of most genomes challenges this assembly process as multiple possibilities for the elongation of contigs often have to be considered (Schatz et al. 2010). The currently most prevalent sequencing technologies generate short reads (~100-250 nt) in large quantities. For *de novo* genome assemblies, different computational approaches were developed and implemented in so-called “assemblers”. These programs are often adapted to the different sequencing technologies or specific properties of the input data. Two main algorithmic concepts “Overlap Layout Consensus” (OLC) and “de Bruijn Graphs” (DBG) will be described in the next paragraphs followed by additional methods to further interconnect contigs to longer units called “scaffolds”.

1.1.3.1 Overlap Layout Consensus

The OLC approach consists of three steps. Overlaps of the input reads are determined by performing all-versus-all alignments. Then, a layout graph is built where nodes represent reads and edges reflect the previously calculated overlaps. A path through this graph is calculated in which each node can only be passed once. Finally, the reads along this path are connected to a consensus sequence. However, the third step can be seen as Hamilton path problem which is known to be NP-complete (Garey and Johnson 1979). The number of nodes is equivalent to the number of reads while the number of edges increases logarithmically with increasing read number (Li et al. 2012). The most popular OLC assemblers are CAP3 (Huang and Madan 1999), Celera Assembler (Myers 2000), Arachne (Batzoglou 2002), Phusion (Mullikin and Ning 2003), Newbler (Margulies et al. 2005) and Phrap (de la Bastide and McCombie 2007). OLC assemblers were initially developed for long read sequencing technologies (Sanger or 454). When high-throughput short read data became available, OLC assemblers became less popular due to poor runtime performance resulting from the increased computational complexity of the overlap calculations and the path construction (Nagarajan and Pop 2013).

1.1.3.2 De Bruijn Graph

Whereas the OLC algorithm follows the intuitive thinking to connect something short (the reads) to something longer (a consensus), DBG follows an anti-intuition paradigm, i.e., prior to assembly reads are split into even shorter sequences. For this, reads are broken into sub-strings of k nucleotides (k -mers) which are used to construct a directed graph in which edges are k -mers and nodes reflect the neighborhood relation of the k -mers. Then, an assembly is derived from an Euler path through this graph where all edges must be used exactly once (Pevzner et al. 2001; Compeau et al. 2011). With increasing read numbers, the number of nodes and edges is nearly constant. Traversing through the

graph is efficiently solvable (in contrast to a Hamilton path) and the nucleotide sequence can be directly inferred without a consensus step. When reads derived from complex genomes with many repeats and/or high heterozygosity are processed with the DBG approach, they often lead to highly fragmented assemblies. Together with sequencing technology improvements and increasing data amounts, DBG assemblers also evolved from the first program called EULER (Pevzner et al. 2001) to more efficient ones like Euler-USR (Chaisson et al. 2009), Velvet (Zerbino and Birney 2008), ABySS (Simpson et al. 2009), SOAPdenovo (Li et al. 2010) or ALLPATHS-LG (Gnerre et al. 2011).

1.1.4 Scaffolding

The process of connecting and orienting contigs is called “scaffolding”. For this, sequencing libraries are produced with a DNA fragment length of longer than twice the read length, which is usually between 0.4 and 40 kb. Both ends of such fragments can be sequenced either directly (“paired-end” libraries, 0.4-1 kb) or after circularization (“mate-pair” libraries, 1-40 kb). As the distance between the two reads is defined by the library preparation protocol, this information is used for scaffolding. In the different library preparation methods mentioned above, read pairs are oriented in different directions, that is, paired-end reads are pointing to one another whereas mate-pair reads point away from the start of circularization. The principle of scaffolding is based on aligning read pairs to a set of contigs and accordingly building connections between contigs and thus establishing a certain order that is supported by read pairs. These connections are usually represented by runs of N (the symbol for an ambiguous nucleotide) whereby the length is estimated from the mean read pair distance. Many of the aforementioned assemblers have incorporated a scaffolding functionality but there are also stand-alone scaffolders like SSPACE (Boetzer et al. 2011), MIP (Salmela et al. 2011), OPERA (Gao et al. 2011) or GRASS (Gritsenko et al. 2012). Often, multiple paired read libraries with different fragment sizes are used for scaffolding, ideally with library sizes adapted to respective genomic properties (Wetzel et al. 2011).

1.1.5 Optical Mapping

Optical mapping provides another type of long range mapping information beneficial for complex genome assemblies (Valouev et al. 2006) and was developed by David C. Schwartz in the 1990s (Schwartz et al. 1993). Long DNA molecules (>250 kb) are immobilized and stretched on a glass surface and subsequently fragmented with a restriction enzyme. The length and sequential arrangement of the resulting digested DNA fragments are then detected as restriction profiles by an optical device. Overlapping restriction profiles are computationally connected to so-called “maptigs”. A set of maptigs results in an optical map for a particular genome (Aston et al. 1999). Sequence contigs or scaffolds are computationally cut (i.e. *in silico* digested) at the specific sites of the restriction enzyme used for experimental DNA digestion producing restriction profiles that are then aligned to the maptigs. Accordingly, the sequence contigs or scaffolds are ordered and oriented, such that longer scaffolds can be formed. This approach can only be successfully applied if the sequence-based assembly contains contigs or scaffolds long enough to show a unique restriction pattern. Optical

mapping was successfully applied to the microbial genomes of *Plasmodium falciparum* (Lai et al. 1999) and *Yersinia pestis* (Zhou et al. 2002) and assisted more recently in the assembly of the maize (Zhou et al. 2009) and goat (Dong et al. 2013) genomes. Currently, the company OpGen (Gaithersburg, Maryland, USA) is commercially providing the described optical mapping technology. The approach of BioNano Genomics (San Diego, California, USA) is comparable but uses additionally nanometer-sized channels to stretch and immobilize the DNA molecules. Instead of digesting DNA molecules at specific restriction sites, the DNA is fluorescently labeled at target sequences and this pattern is optically detected and can be interpreted and analyzed similarly to OpGen's approach.

1.1.6 Genetic Mapping

In contrast to sequence-based and optical mapping-based scaffolding, for which physical distances are used, genetic map-assisted scaffolding makes use of genetic distances. A genetic or linkage map shows the genetic distance between markers and their order relative to each other, and is organized in so-called linkage groups (LGs); at minimum a linkage group includes three markers. Markers are considered linked if they are located on the same chromosome. The linkage-based order of unique markers (sequences) is employed for assessing existing scaffolds as well as for extending them, and for connecting and ordering previously separate scaffolds.

The underlying biological process for building a linkage map is the homologous recombination of chromosomes during meiosis called crossing-over. To build a linkage map, unique genomic sequences (markers) are required that need to be polymorphic and should be evenly distributed in the genome under investigation. Markers can include single nucleotide polymorphisms (SNPs) (Cho et al. 1999), microsatellites (Weissenbach et al. 1992) or other sequence variations like for example restriction fragment length polymorphisms (Botstein et al. 1980). The distance between two markers is calculated from the frequency with which they get disconnected during crossing-over. The probability for markers to get disconnected is higher if the physical distance between two markers is longer. The genetic distance between two markers is given in centi Morgan (cM) where 1 cM represents the probability to recombine once in 100 meioses. Both order and distances of markers are calculated resulting in a linkage map that usually contains multiple LGs: Loosely linked or un-linked marker groups are assumed to be located on different chromosomes. The method of anchoring sequences based on genetic and other maps was applied in the human genome sequencing project (Lander et al. 2001), in which different high-resolution linkage maps with either more than 5,000 (Dib et al. 1996) or even 8,000 (Broman et al. 1998) microsatellite markers were used. More recently, genetic maps were constructed and employed in the genome assemblies for example of watermelon (Ren et al. 2012) or potato (Sharma et al. 2013).

1.1.7 Synteny Scaffolding

The term “synteny” is controversially discussed (Passarge et al. 1999), so a definition for the purpose of this work is needed. Synteny refers to genes, or more general to sequences in a genome, which are located on the same chromosome (Nadeau 1989). Further, the term “conserved synteny” refers to homologous genes that are syntenic in two or more species, regardless of their order on these chromosomes (Ehrlich et al. 1997). In addition to conserved synteny a “conserved order” can be utilized to improve genome assemblies. Blocks of genes with conserved order can be identified by pair-wise alignments of coding sequences (CDSs) of two species. In genome assembly, scaffolds can be tentatively assigned to the same chromosome based on conserved synteny and potentially arranged and oriented according to their conserved gene order on a chromosome of a closely related species. This approach recently supported the anchoring of goat scaffolds onto chromosomes using synteny comparisons to the cattle genome (Dong et al. 2013). The ancestral karyotype of medaka and human was also reconstructed by using synteny analyses (Kasahara et al. 2007).

1.2 Repetitive Elements in Genomes

The complexity of genomes is not strictly correlated with its gene content but rather with the abundance of repetitive elements. Even in relatively compact microbial genomes, 1-6% is formed by repetitive sequences (Koressaar and Remm 2012) but especially in large genomes of higher organisms the fraction of such repeats can exceed its unique fraction. For example, in the pufferfish *Tetraodon nigroviridis* repeats make up 6% of the genome (Crollius 2000), in the fruit fly *Drosophila melanogaster* 12% (Sackton et al. 2009), in human 45% (Lander et al. 2001) and in barley 84% (International Barley Genome Sequencing Consortium 2012). In the early days of genomic research, the functions and characteristics of repeats were not well understood and the term “junk DNA” was broadly associated with sequences whose function is not known (Ohno 1972). Later, the concept of selfish or parasite DNA was introduced. Selfish DNA spreads in the genome by forming additional copies of it but its presence has little or no effect on the phenotype (Orgel and Crick 1980). In recent years, more and more genomes were analyzed leading to a significant increase of knowledge about repetitive elements. It is now recognized that repeats are involved in reshaping genomes by several mechanisms like generating insertion mutations and genomic instability, by altering gene expression or by contributing to genetic innovation (Feschotte and Pritham 2007; Cordaux and Batzer 2009). Due to the large diversity of repeats, these elements can be categorized based on several factors like size, frequency of occurrence, distribution pattern, biological function, or replication mechanism. A widely accepted classification scheme is outlined in the following sections.

1.2.1 Tandem Repeats

Tandem repeats (TRs) are composed of highly conserved sequence motifs located directly adjacent to each other. Based on the length of these repeating motifs, TRs are categorized into (i) microsatellites (one unit comprised of 1 to 5 nt), (ii) minisatellites (6-99 nt) or (iii) satellites (>100 nt) (Lim et al. 2013). TRs up to the unit size of 6 nt are also called simple sequence repeats (SSRs) or short tandem

repeats (STRs) (Merkel and Gemmell 2008). Mechanisms that lead to the accumulation of TRs in a genome are errors in DNA replication (replication slippage) and unequal crossing-over. Microsatellites mainly occur as di- or tri-nucleotide repeats and have a much higher mutation rate compared to other genome regions (Gemayel et al. 2010). The elevated mutation rate allows the study of genetic variability of species or populations. TRs are also involved in a number of diseases. If a microsatellite is located in a protein coding gene, its mutation can lead to loss-of-function or gain-of-function of that gene, as for Huntington disease or the fragile X syndrome in humans (Gemayel et al. 2010). In contrast to micro- and minisatellites, satellites are typically organized in longer arrays that can occupy several million nucleotides. The probably best-described satellite is the 171 nt human alpha-satellite which is predominantly located in centromeric and peri-centromeric regions of all human chromosomes (Rudd et al. 2006). Alpha-satellites play a key role in the genome replication process because kinetochore proteins attach to these large homogeneous arrays (Erliandri et al. 2015).

1.2.2 Transposable Elements

Transposable elements (TEs) are found in the genomes of almost all living species. Because they have the ability to self-replicate and move in a genome, it is not unusual that a major fraction of the DNA is formed by TEs. In plants, they can populate up to 80% of the genome and in vertebrates almost 50%. These elements are categorized based on their replication mechanism and enzymatic properties into retrotransposons (often and also herein referred to as “Class I”) and DNA transposons (“Class II”) (Finnegan 1989; Wicker et al. 2007). With a few exceptions, Class I elements are more abundant in eukaryotic genomes whereas Class II elements predominate in bacteria. Most elements of both classes share a feature called Target Site Duplication (TSD), resulting from filling of staggered nicks generated at the DNA target site upon insertion of TEs (Lopez-Flores and Garrido-Ramos 2012).

1.2.2.1 Class I - Retrotransposons

During the so-called “copy-and-paste replication” of Class I elements, an RNA intermediate is built from the DNA template and is incorporated as copy of the template into a new location of the genome (Boeke and Corces 1989). This process is mediated by polyproteins that are encoded in one to three CDSs of the elements’ sequence. Five orders of retrotransposons can be discerned: (i) Long terminal repeat (LTR) retrotransposons, (ii) Dictyostelium Intermediate Repeat Sequences (DIRSs), (iii) Penelope-Like Elements (PLEs), (iv) Long and (v) Short Interspersed Nuclear Elements (LINEs / SINEs). LTR retrotransposons are flanked by LTRs at both ends that carry promoter sequences necessary for the transcription of the element. At the integration site, a 4-6 bp long TSD is generated. LTR retrotransposons are common in plants where they make up the largest fraction of repeats whereas they are less frequent in animals. DIRSs were first described in the amoeba *Dictyostelium discoideum* but were also found in animals, plants and fungi. DIRS share many features of LTR retrotransposons but they are using a different protein for their integration process. Therefore, the flanks of DIRSs do not carry TSDs or LTRs but rather terminal inverted repeats (TIRs) (Wicker et al. 2007). PLEs are present in unicellular animals, fungi and plants but are not as widely distributed as

LTR retrotransposons (Evgen'ev and Arkhipova 2005). PLEs have only one CDS coding for a reverse transcriptase (RT) which differs in its sequence from that of other Class I elements. In some cases, the RT gene contains an intron, which is retained after retrotransposition. PLEs build relatively long TSDs of 10-15 nt at their flanks upon integration (Jurka et al. 2007). LINEs and SINEs are not flanked by LTRs but build TSDs of variable lengths. LINEs encode one or two retrotransposition proteins whereas SINEs lack these proteins but rather use those encoded by LINEs. SINEs emerged *de novo* several times in evolution from for example transfer RNA (tRNA) or ribosomal RNA (rRNA) molecules (Kramerov and Vassetzky 2011). In humans, the most frequent and best studied LINEs are members of the L1 family. They have a size of 6-8 kb and occur in up to 516,000 copies in the genome. The most frequent SINEs are members of the Alu family. Alus are only 300 nt long and occur 1.1 million times in human. Both element types together account for 34% of the human genome (Lander et al. 2001).

1.2.2.2 Class II - DNA Transposons

In contrast to retrotransposons, Class II elements can move without any RNA intermediate (Craig 1995). Therefore, DNA transposons are often less abundant than other TEs, for example in humans they comprise only 3% of the genome (Lander et al. 2001). Three types of Class II elements are known that are (i) “cut-and-paste” transposons, (ii) Helitrons and (iii) Polintons. Cut-and-paste transposons are the most prevalent elements and excise their double stranded DNA and re-insert it into a new location, thereby creating a TIR at the ends. To facilitate transposition, the transposase protein of the element recognizes the TIRs and cuts both strands of the DNA and re-inserts it at the target position. In contrast, Helitrons and Polintons (also called Mavericks) cut the DNA only at one strand at each end and transfer their sequence by a single-stranded DNA to the target position, which results in a duplication of the element rather than a movement (Wicker et al. 2007). Helitrons are moving via a “rolling-circle” mechanism without showing TSDs or TIRs but having a characteristic hairpin motif at the ends (Kapitonov and Jurka 2007). Polintons can reach sizes of 10-20 kb with having long TIRs (100-1,000 bp) and can code for multiple proteins. These are a DNA polymerase and a retroviral-like integrase, which is a typical feature of Class I elements. In contrast to those, Polintons lack an RT domain, which hinders them from forming an RNA intermediate and therefore categorizes them as Class II elements (Feschotte and Pritham 2007). Some Polintons can also code for capsid proteins which in principle allow the formation of virion particles (Krupovic and Koonin 2014).

1.2.3 Gene Families and Segmental Duplications

A gene family is a group of genes that arose from a duplication, e.g. whole genome or local duplication. For example, in many eukaryotes, rRNA genes coding for the subunit proteins 28S, 18S, 5.8S and 5S are arranged in tandem and present at multiple genomic loci (Long and Dawid 1980). In humans, well known examples include the histone genes H1, H2A, H2B, H3, and H4, which are organized mainly in two clusters on chromosomes 1 and 6 (Albig et al. 1997), the genes of the major histocompatibility complex on chromosome 6 (MHC Sequencing Consortium 1999), the

immunoglobulin genes on chromosomes 2, 14, 22 (Honjo 1983) and olfactory receptor genes, which are distributed over several chromosomes with the highest abundance on chromosome 11 (Niimura and Nei 2003).

Segmental duplications (SDs) are highly similar and relatively long regions often found in large genomes. They are formally defined as being at least 90% identical and longer than 1 kb (Bailey et al. 2001). SDs can contain genes, common repeats or sequences with unknown function. The human genome is comprised of 5% SDs which occur inter- as well as intrachromosomally (Bailey and Eichler 2006).

1.2.4 Identification of Repetitive Elements

Repeats present major challenges in sequence analyses such as read alignment, *de novo* genome assembly and genome annotation (Treangen and Salzberg 2012). Therefore, identification and classification of repeats are among the first steps when genome assemblies are created or annotated, as the CDSs of transposons can be mistakenly interpreted as non-repetitive genes (Tang 2007; Yandell and Ence 2012). Repeats are identified by using different approaches. Selecting the most appropriate one depends on several factors: (i) overall repeat content of the genome of interest, (ii) amount and type of available sequence data, (iii) previous knowledge like repeat databases, and (iv) available computational resources.

If no sequence information is available, repeats can be analyzed experimentally in the laboratory, but if there are sequence resources known, they can be analyzed with either similarity-based or *ab initio* methods. These computational approaches are applicable either to unassembled reads or to an assembled genome. Many programs are available and were evaluated in several reviews (Bergman and Quesneville 2007; Saha et al. 2008a; Lerat 2010). In the following, I will shortly summarize the most popular programs including those relevant for this thesis.

1.2.4.1 Experimental Repeat Identification

If sequence information is lacking, one can detect repetitive DNA experimentally. For this, double-stranded DNA is fragmented and denatured at high temperature. Next, the temperature is lowered again and thus single strands re-associate to form double-stranded molecules whereby abundant fragments such as repeats hybridize faster than those of the unique fraction. Monitoring this process over time allows the estimation of the repeat content of a genome (Britten and Kohne 1968). This approach was employed for obtaining the initial estimate of presence and number of Alu elements in the human genome (Houck et al. 1979). Other methods like electron microscopy identified hairpin-like structures resulting from self-hybridization of the palindromic Alu sequence (Deininger and Schmid 1976). In saturation hybridization experiments, the abundance of L1 elements in the mouse genome was estimated. For this, DNA molecules carrying L1 sequences are immobilized on filters and the percentages of radioactive total DNA retained by the filters is determined (Gebhard et al. 1982). The structure of L1 elements and their abundance in mammalian species was determined by Southern

blotting, where characteristic restriction patterns can identify subparts of the repeat elements and therefore reveal their approximate internal organization (Burton et al. 1986).

1.2.4.2 Similarity-Based Repeat Identification

This computational method analyses the input sequence data for sequence homology to a set of already known repeat consensus sequences, commonly called “repeat library”. Often, genome sequencing projects provide repeat libraries, which are specific for a particular species but there are also repeat databases that collect and classify well known as well as newly discovered repeat consensus sequences.

RepBase Update (short RepBase) (Jurka 2000; Jurka et al. 2005) is the largest and most comprehensive repeat database containing over 46,000 (version 20.09) elements from a large variety of eukaryotic species and has been constantly growing since its foundation in 1992. Other databases are The TIGR Plant Repeat Databases (Ouyang and Buell 2004) or the Triticeae repeat database TREP (Wicker et al. 2002), which are both more focused on plant genomes.

The most popular programs for sequence similarity-based repeat identification are RepeatMasker (Smit et al. 1996-2015) and Censor (Jurka et al. 1996; Kohany et al. 2006). Both use by default RepBase but can also integrate an individually created repeat library. The sequence homology searches done by RepeatMasker can be performed by different alignment programs such as crossmatch (Green and Ewing 2002) or different BLAST derivatives. RepeatMasker finally provides statistics for the identified repeats as well as a revised version of the input sequence in which repetitive nucleotides are changed to ambiguous bases, i.e. masked with “N”. The program Censor was developed by the RepBase curators and authors. It also compares input sequences to a repeat library, masks repetitive regions and reports repeat classifications.

1.2.4.3 De novo Repeat Identification

The *de novo* detection of TRs is usually done prior to the detection of DRs. In general, two computational steps are included: (i) detection and (ii) filtering (Lim et al. 2013). Detection can be done by either a combinatorial or a statistical/heuristic approach. A subsequent filtering step evaluates the candidate TRs and discards false positives. Many tools for TR detection are available, and some have benefits over others in respect to specificity or run time. A widely-used program is Tandem Repeats Finder (TRF, Benson 1999). TRF is fast, accurate and detects TRs with a unit size of up to 2 kb (Lim et al. 2013).

To *de novo* identify dispersed repeats (DRs), programs such as RECON (Bao and Eddy 2002), RepeatScout (Price et al. 2005) or ReAS (Li et al. 2005) are used. These directly exploit the repetitiveness of these elements without any prior knowledge (Saha et al. 2008b). RECON is one representative of a group of programs that perform self-comparison alignments of the input sequences. Regions with a high degree of overlap are assumed to belong to the same repeat type and consensus

sequences are built for each type. Another type of programs like RepeatScout or ReAS, reduce in a first step the search space to short sub-strings (k-mers). The frequency of the k-mers is determined and only those above a certain threshold are aligned to the input sequences and subsequently extended to form a consensus. In contrast to RepeatScout, which only works with assembled genomic sequences, ReAS performs the repeat identification directly on sequence reads and was initially designed to work with Sanger sequencing reads. This raw-data strategy circumvents errors potentially introduced by an incorrect assembly (Saha et al. 2008b) and can identify repeats that are not present in a given assembly (Li et al. 2005). Only little effort to pursue the read-based approach was made, so only few programs were developed for analyzing 2nd generation short read data. One of those is RepeatExplorer (Novak et al. 2013), which is incorporated in the Galaxy framework (Goecks et al. 2010) and analyzes similarities between NGS reads to build graphs that represent repeat families. The program Transposome uses and improves that graph-based approach (Staton and Burke 2015).

1.2.4.4 Signature-Based Repeat Identification

Because several classes of TEs contain structural features or alter the host genome at their insertion sites in a special way, one can as well search for such patterns. For example, the program LTRharvest (Ellinghaus et al. 2008) finds new LTR retrotransposons based on features like LTR size range, LTR distance and TSD presence. RTAnalyzer (Lucier et al. 2007) detects signatures of L1 retrotransposons like a TSD, a polyA tail or an endonuclease cleavage site. Many more programs have been designed, also for other TE types like Helitrons or Miniature Inverted Repeats (MITEs), and they can be efficient for particular applications. However, signature detection is limited by the knowledge of structural features of classified TEs and cannot detect repeats without conserved features or with novel structures (Lerat 2010).

1.2.4.5 Integrated Repeat Identification and Annotation Pipelines

Due to the broad variety of repeat analysis tools, which have advantages and drawbacks, automated pipelines were developed (Lerat 2010). REPET is such a package, and combines two modules: TEdenovo and TEannot (Flutre et al. 2011). TEdenovo employs BLASTER, Grouper (both programs: Quesneville et al. 2003), RECON and PILER (Edgar and Myers 2005) to identify and cluster repeats which are then classified with an existing repeat library by TEannot. RepeatModeler is another pipeline (Smit and Hubley 2008-2010) that identifies and classifies repeats using a combination of RepeatMasker, RECON, RepeatScout and TRF. Here, the output of the *de novo* repeat identification programs RECON and RepeatScout is used to build, refine and classify consensus sequences of TEs.

1.2.4.6 Repeat Classification Programs

The number of programs designed to assign a class or family to a repeat consensus is much smaller than the number of repeat identification programs. One reason for this is that the classification scheme of repeats is constantly developing, as new types of elements and relationships between elements are

being discovered (Wicker 2007). The most recent programs developed for the automated classification of an uncharacterized repeat library are TEclass (Abrusan et al. 2009) and REPCLASS (Feschotte et al. 2009), which are based on different working principles. TEclass aims to classify repeats according to the main classes of elements and is based on a machine learning approach that compares oligomer frequencies of already classified repeats (for example from RepBase) to the query sequences. REPCLASS consists of three independent modules: One compares homologies to known repeats, the second searches for structural features of different classes of repeats and the third identifies TSDs. The results of the different modules are finally combined. However, in many cases a large fraction of repeats remain unclassified due to possible incompleteness of the query sequences, high divergence to known repeats or just because they are novel and not yet described.

1.3 *Nothobranchius furzeri*

Biological processes are often studied in model organisms. Several properties are relevant that designate a particular organism as a model, such as housing and breeding conditions, physical suitability or the evolutionary relation to other species to which the respective results are to be subsequently transferred. Aging research aims for understanding the biological mechanisms underlying aging. For this, an appropriate model organism should on the one hand be as closely related to humans as possible but on the other should allow the study of aging processes in a practical time frame. Several short-lived animal models were established for this research field like *D. melanogaster* or the worm *Cenorhabditis elegans* but they are invertebrates and thus phylogenetically far away from humans. The Turquoise Killifish (*Nothobranchius furzeri*) is a short-living vertebrate species and therefore well suitable to study aging in a more closely related and complex organism.

1.3.1 *N. furzeri* is a Model Organism in Aging Research

N. furzeri is a small freshwater species from the South-East of Africa. It lives in ponds found in a region with large differences in the amount of precipitation over the year. Most of the rain falls in a short period of time, i.e. in the rainy season, in which ponds are formed or filled up. The fish have adapted to the regular disappearance of the habitat by developing a specialized life cycle that is characterized by fast development and reaching sexual maturity within few weeks. At the end of the rainy season, the parent generation disappears together with the water, whereas the embryos survive the dry season in a dormant state (called diapause), encased in the dry mud. Embryos hatch when the water comes back, and thus a new generation grows up (Levels et al. 1986).

Various *N. furzeri* populations originating from different locations are known. The very first *N. furzeri* specimens were caught in 1968 in the Gonarezhou National Park in the south of Zimbabwe by Furzer and Warne (Jubb 1971) and their subsequent breeding in the laboratory resulted in the current strain named “GRZ” (Valdesalici and Cellerino 2003). GRZ fish were kept by hobby aquarists and descendants are still available for research. Another laboratory strain stems from a population from Mozambique, 300 km south of the GRZ collection site and is called “MZM-0403” (Terzibasi et

al. 2008). The two collection sites differ in humidity. The GRZ locality in Zimbabwe receives on average 300 mm precipitation per year whereas for the other locality in Mozambique 600 mm per year are observed. More strains from other locations are available and a systematic analysis of newly collected populations was published recently (Bartakova et al. 2013). These natural populations and their descendant strains show different lifespans. GRZ fish show a median lifespan (50% survivors) of 10 weeks and a maximum lifespan (10% survivors) of 16 weeks whereas MZM-0403 have a median and maximum lifespan of 24 and 31 weeks (Terzibasi et al. 2008; Hartmann et al. 2009). When strains with differing lifespans are inter-crossed, their F1 offspring shows an intermediate lifespan (Kirschner et al. 2012). It was also found that the median and maximum lifespan of *N. furzeri* increases upon administration of the antioxidant resveratrol (Valenzano et al. 2006b), a reduced water temperature (Valenzano et al. 2006a) and dietary restriction (Terzibasi et al. 2009). Moreover, *N. furzeri* shows a clearly visible and measurable aging phenotype. Finally, inbred lines (GRZ) are available. These properties qualify *N. furzeri* as suitable model in aging research (reviewed in Genade et al. 2005; Scharl 2014; Cellerino et al. 2015).

1.3.2 Genetic and Genomic Resources

To fully employ *N. furzeri* as a model for aging research, comprehensive genetic and genomic resources are essential. A number of sequence data as well as genetic markers and maps have been available at the beginning of this thesis, which will be summarized in the following.

1.3.2.1 Initial Characterization of the *N. furzeri* Genome

In an initial characterization of the genome, two genome size estimates were carried out. One was based on assessing the gene content in a 5.4 Mb sample of Sanger sequences (1.6 to 1.9 Gb) whereas the other utilized flow cytometry measurements (1.5 Gb) (Reichwald et al. 2009). The *N. furzeri* genome is diploid with 19 chromosomes ($2n=38$) and lacks morphologically discernible sex chromosomes. One distinctive feature of the genome is its high repeat content, which is assumed to contribute to the comparatively big genome size. For the sequence sample comprising 5.4 Mb, repeats were estimated at 45%, composed of 25% DRs and almost 21% TRs. The amount of TRs is exceptionally high compared to four fish species that were analyzed for comparison (medaka, stickleback, tetraodon and zebrafish), in which TR content only ranges from 1.7% in medaka to 5% in zebra fish. In *N. furzeri*, a minisatellite with a unit size of 77 bp and a satellite with a unit size of 348 bp were found to represent the main fraction of the TR content. Their genomic location is in the centromeric and peri-centromeric region of the chromosomes, as found by fluorescence *in situ* hybridization (FISH) (Reichwald et al. 2009). A second analysis of an extended Sanger sequence dataset (120 Mb) supported the high fraction of TRs and identified additional DRs, thus resulting in a total repeat content of 64% in the genome of *N. furzeri* (Koch 2010).

In the initial work, also the genetic variation in the GRZ and the MZM-0403 *N. furzeri* strains was assessed. It was shown by genotyping gene-associated single nucleotide variations (SNVs) and

microsatellite markers that the GRZ strain is highly inbred while the more recently collected MZM-0403 strain still resembles the variability of the wild population (Reichwald et al. 2009).

1.3.2.2 Linkage Maps of the *N. furzeri* Genome

In 2009 and 2012, genome-wide genetic linkage maps were constructed, both based on *N. furzeri* intra-species crosses of GRZ and MZM-0403 specimens. In the first cross, 413 F₂-individuals were genotyped for 152 microsatellites, and a linkage map comprising 25 linkage groups was constructed (Valenzano et al. 2009). Additionally recorded phenotypic data allowed the identification of regions in the linkage map that were associated with sex determination and tail color. This work also found the sex determining system to be XX/XY with the male sex being the heterogametic sex in *N. furzeri* (Valenzano et al. 2009). In a second cross, 404 F₂-individuals were genotyped for 411 marker loci of which 283 are gene-associated SNVs and 128 are microsatellites. The resulting linkage map contained 22 LGs, of which three likely represented fragments of the other LGs. The number of remaining 19 LGs agreed well with the number of chromosomes determined by karyotyping, and the total length of the map was in accordance to the estimated size of the genome (Kirschner et al. 2012). A third map was constructed from an inter-species cross between a *N. furzeri* GRZ male and a female of the sister species *N. kadleci* resulting in 287 F₂ offspring. The related genetic map contains 237 SNV markers in 20 LGs (Ng'oma et al. 2014).

1.3.2.3 Sequencing Resources

For the genome assembly and other analyses like variation detection or phylogenetic studies, a large amount of data was generated. At the time when I started working on this thesis, 120 Mb Sanger sequences, 8 Gb of 454, and 166.4 Gb of Illumina reads were available. As an additional, independent data source, RNA from individuals of different *N. furzeri* strains, of multiple tissues and time points was sequenced for transcriptome assembly (Petzold et al. 2013) and for performing expression studies in the aging process (JenAge⁷).

⁷ <http://www.jenage.de>

1.4 Thesis Objectives

Aim 1: Over the last few years, *N. furzeri* has been established as a new vertebrate model organism for the studies of vertebrate aging, which is based on its exceptionally short lifespan and the presence of typical aging-related characteristics (Genade et al. 2005). One prerequisite for a model organism is the availability of comprehensive genetic and genomic resources. Although a transcript catalogue was established, a lot of genomic information is still undiscovered. By sequencing, assembling and annotating the *N. furzeri* genome, an important resource will be provided to the research community.

Approach: To this end, the *N. furzeri* genome is sequenced using the Sanger, 454, Illumina and PacBio technologies. In addition, resources from optical and genetic mapping as well as synteny comparisons to other species are incorporated. Bioinformatics tools and approaches for the processing of data and assembly of the genome are developed and applied. In a multi-step process, the assembly is continuously improved to obtain a high-quality genome reference sequence on a chromosomal scale.

Aim 2: Complex genomes, like that of *N. furzeri*, are composed of many repeats. Their detection is challenging and currently available methods were mainly designed for Sanger-like or already assembled data and are not applicable to high-throughput raw datasets. Therefore the second aim of this thesis is to develop a method that detects repetitive elements in 2nd generation datasets.

Approach: A software pipeline is established that analyzes k-mer frequencies in Illumina datasets and determines whether they are repeat-derived or belong to the unique fraction of the genome. These high frequent k-mers are assembled in a library of consensus sequences representing the repetitive elements of the analyzed species. The pipeline incorporates established and efficient programs (Jellyfish, Velvet) and is evaluated on the *D. melanogaster* and human genome.

Aim 3: The *N. furzeri* genome is featured by an exceptionally high repeat content (Reichwald et al. 2009). To support the genome assembly as well as for further genome-wide analyses, these repetitive elements need to be discovered and characterized. In particular, a most comprehensive repeat library will allow a more precise estimation of the composition of the *N. furzeri* genome, a thorough repeat annotation of the assembly and a repeat-base completeness assessment of the reference sequence.

Approach: Using different genomic datasets of *N. furzeri*, three independent methods are applied to identify repeats. A (i) genome assembly-based (RepeatModeler), a (ii) Sanger reads-based (RepeatScout) and a (iii) Illumina reads-based (RepARK) analysis are performed and the results are combined into a *N. furzeri*-specific repeat library. This library is then used to characterize the repeat content of the genome, the reference sequence and the yet unassembled fraction.

2 Material and Methods

2.1 Assembly of the *N. furzeri* Genome

2.1.1 DNA Sequencing

Four sequencing technologies were applied to sequence the DNA of eleven individuals of the GRZ strain (Table 1). DNA was isolated from skin, muscle, or whole body with the DNeasy Mini/Midi kit (Qiagen) and sequenced either on the Applied Biosystems 3730xl DNA Analyzer, Illumina Genome Analyzer IIx, Illumina HiSeq 2000/2500, Roche 454 GS FLX Titanium or Pacific Biosciences RS II instrument.

Table 1: *N. furzeri* GRZ Individuals in this Study and Their Use for Various Analyses.

Individual	Sex	Sequencing Technology	Analyses
fish1	male	454 & Illumina	KILAPE scaffolding, GapFiller, GapCloser
fish2	male	454	repeat analysis
fish3	male	454, Illumina & Sanger (WGS)	KILAPE scaffolding, GapFiller, GapCloser, repeat library construction
fish31	female	Illumina	ALLPATHS-LG assembly, GapFiller, GapCloser, repeat library construction (RepARK)
fish55	female	Illumina	ALLPATHS-LG assembly
fish1a	male	Sanger (BAC ends)	Scaffold placement after optical map integration
fish3a	male	Sanger (BAC ends)	Scaffold placement after optical map integration
fish5a	male	Sanger (BAC ends)	Scaffold placement after optical map integration
M004	male	Illumina	repeat analysis
M005	male	Illumina	repeat analysis
M013	male	PacBio	Gap filling, repeat analysis, assembly completeness

2.1.1.1 Sanger Sequencing

Genomic DNA of the GRZ male fish3 was extracted and sequenced as described in Reichwald et al. (Reichwald et al. 2009), resulting in 132,390 sequences comprising 120 Mb (Table 2). Genomic DNA of male *N. furzeri* GRZ specimens fish1a, fish3a and fish5a was used to make a bacterial artificial chromosome (BAC) library comprising 129,024 clones (done by the Clemson University Genomics and Computational Lab – GCL, Clemson, South Carolina, USA). The average length of *N. furzeri* BAC inserts is 145-150 kb (estimated by GCL). Direct Sanger sequencing of BAC insert ends resulted in 108,994 sequences (Table 2).

2.1.1.2 454 Sequencing

For long-range scaffolding as well as repeat analyses, 3 kb, 8 kb and 20 kb libraries were prepared from male specimens fish1, fish2 and fish3 and sequenced on the 454 GS FLX platform. This resulted in 62 mate-pair sequencing datasets from 3 kb libraries, 21 datasets from 8 kb libraries and 17 datasets from 20 kb libraries with a total of 17.6 Gb (Table 2).

2.1.1.3 *Illumina Sequencing*

For the initial genome assembly, 12 DNA libraries derived from two female specimens, fish31 and fish55, were sequenced yielding 305 Gb of raw data (Table 3). Two libraries with an insert size of 170 bp (fish31) and ten libraries with an insert size of 3 kb (fish55) were sequenced with the Illumina HiSeq2000 in 2x100 bp paired-end or mate-pair mode, respectively.

Additionally, seven 300 bp-insert libraries prepared from male GRZ specimens fish1, fish3, M004, and M005 as well as one 3 kb-insert library derived from fish3 were sequenced with the Illumina HiSeq 2000/2500 in 2x100 bp paired-end mode for the 300 bp libraries or mate-pair mode for the 3 kb library. Sequencing resulted in 156.6 Gb for the 300 bp libraries and 34.7 Gb for the 3 kb library (Table 2).

2.1.1.4 *PacBio Sequencing*

Genomic DNA of the male specimen M013 was isolated and fragmented. Fragments were converted into sequencing libraries, size-selected for a range of 10-50 kb and sequenced on a PacBio RS II system resulting in 453,425 subreads comprising 2.46 Gb (Table 2). Using these subreads, 10,987 (27.1 Mb) so-called “reads of insert” (ROIs) (Table 2) were build with the protocol RS_ReadsOfInsert_v1 from the SMRT Analysis package (v.2.3.0)

Table 2: Sequence Data Used for Scaffolding with KILAPE, Gap Filling and Repeat Analyses.

ID	Accession	Specimen	Pairing type ^d	Insert size	Reads	Bases	Coverage ^a
Sanger							
wgs.FLI_Nfu_GRZ ^b	unpublished ^c	fish3	-	-	132,390	120,394,892	0.06
bac.FLI_Nfu_GRZ	LIBGSS_039197	fish1a, fish3a, fish5a	clone ends	150 kb	108,994	81,339,503	0.04
454							
gDNA_fish1 pe8	ERR754559- ERR754564	fish1	mate-pair	8 kb	4,190,549	1,344,792,695	0.7
gDNA_fish3 pe8	ERR754565- ERR754579	fish3	mate-pair	8 kb	8,982,703	2,561,271,252	1.4
gDNA_fish3 pe20	ERR754580- ERR754582 & ERR754584- ERR754597	fish3	mate-pair	20 kb	7,838,770	2,205,200,292	1.2
gDNA_fish3 pe3	ERR754516- ERR754558	fish3	mate-pair	3 kb	26,027,461	7,491,345,476	3.9
gDNA_fish2 pe3	ERR754497- ERR754515	fish2	mate-pair	3 kb	13,243,670	3,959,959,288	2.1
Illumina							
100414	ERR583466	fish1	paired-end	300 bp	66,979,980	6,764,977,980	3.6
100512	ERR583467	fish1	paired-end	300 bp	392,898,118	39,682,709,918	20.9
110118_mp	ERR583469	fish3	mate-pair	3 kb	353,449,569	34,733,225,706	18.3
110118_pe	ERR583468	fish3	paired-end	300 bp	683,033,726	68,986,406,326	36.3
140826	ERR983246	M004	paired-end	300 bp	62,583,572	6,320,940,772	3.3
140826	ERR983247	M005	paired-end	300 bp	72,096,848	7,281,781,648	3.8
140917	ERR983248	M004	paired-end	300 bp	125,424,906	12,667,915,506	6.7
140917	ERR983249	M005	paired-end	300 bp	147,525,874	14,900,113,274	7.8
PacBio WGS subreads							
wgs.PacBio_GRZ	ERR982683- ERR982688	M013	-	-	453,425	2,462,655,510	- ^d
wgs.PB_ROI_GRZ	unpublished	M013	-	-	10,987	27,110,410	0.01

^a Genomic coverage depth calculations are based on 1.9Gb genome size. ^b WGS reads are a collection of contigs where overlapping ends were joined and of single reads of either end of a fragment. ^c The published equivalent of this dataset (125,133 sequences; 118.6 Mb) was filtered according to submission criteria and can be accessed under ABLO00000000. ^d A coverage value based on subreads is misleading and is therefore not shown.

Table 3: Sequence Data Used for the ALLPATHS-LG Initial Assembly.

Date ID	Accession	Specimen	Insert size [bp]	Insert size SD [bp]	Reads	Bases	Coverage ^a
Paired-end libraries							
110404 ^b	ERR583470	fish31	170	15	706,442,950	70,644,295,000	37.2
110421	ERR583471	fish31	170	15	876,575,866	87,657,586,600	46.1
Paired-end total					1,583,018,816	158,301,881,600	83.3
Mate-pair libraries							
110506	ERR583472	fish55	3,075	403	138,178,629	12,733,092,144	6.7
110627	ERR583473	fish55	3,081	413	50,819,424	4,655,692,137	2.5
110627	ERR583474	fish55	3,098	391	444,284,976	41,944,777,095	22.1
110627	ERR583475	fish55	2,874	379	109,018,480	10,095,623,805	5.3
110627	ERR583476	fish55	2,864	379	49,033,411	4,577,513,117	2.4
110916	ERR583477	fish55	3,468	513	97,094,550	9,304,849,031	4.9
110916	ERR583478	fish55	3,423	554	118,651,237	11,359,094,542	6.0
110916	ERR583479	fish55	2,797	366	152,385,483	14,643,525,393	7.7
111103	ERR583480	fish55	2,554	1,012	198,195,873	18,831,237,521	9.9
111103	ERR583481	fish55	2,510	982	197,496,815	18,450,928,362	9.7
Mate-pair total					1,555,158,878	146,596,333,147	77.2

^a Coverage calculations are based on 1.9 Gb genome size. ^b This dataset was additionally used for gap filling, SD: standard deviation.

2.1.2 DNA Preparation for Optical Mapping

Primary skin fibroblasts from one GRZ female were cultivated at the Leibniz-Institute for Aging - Fritz-Lipmann-Institute (FLI) as previously described (Graf et al. 2013). From the culture, 2×10^7 cells were prepared, frozen at minus 80°C and sent to OpGen Inc. where the extraction of the high-molecular weight DNA, i.e. molecules larger than 250 kb, was performed.

2.1.3 Data Preparation for ALLPATHS-LG

To provide ALLPATHS-LG with an optimal data input, reads of the different libraries of the female specimens fish31 and fish55 were pre-processed as follows: Each mate-pair library was analyzed for duplicate reads, which are assumed to be derived from the same DNA fragment (duplicon). Duplicon detection was done with an in-house Perl script that compares the first 16 nucleotides of the first read and the last 16 nucleotides of the second read against those of all other read pairs. If both 16-mers are identical to those of another mate-pair, they are marked as duplicons. The mate-pair libraries were filtered so that not more than 30 duplicons of a particular fragment remain in the dataset. Of these mate-pair reads, only true pairs, i.e. those of which the forward and reverse read passed filtering, were kept while the paired-end reads were randomly down-sampled according to the recommendations of ALLPATHS-LG (Gnerre et al. 2011).

2.1.4 Basic Assembly with ALLPATHS-LG

The initial *N. furzeri* genome assembly was built with ALLPATHS-LG (version 42316). The program was run on a Supermicro X80BN platform with 64 logical cores and 1,024 gigabytes (GB) of memory using the pre-processed paired-end and mate-pair libraries as input and “-ploidy 2”.

2.1.5 Further Scaffolding with KILAPE and Additional Gap Filling

To improve the contiguity of the assembly, additional scaffolding was done with KILAPE using the 454 mate-pair data as input. KILAPE⁸ is an in-house-developed scaffolding pipeline designed for complex and repeat-rich genomes (B. Downie, personal communication). It predicts and eliminates repetitive motifs directly from high-throughput sequencing read libraries without resorting to a repeat library or a genome reference sequence. Thereby, it aims to minimize scaffolding errors due to repeats. For this, it first identifies and masks erroneous and frequent k-mers in the read datasets. The masked reads are mapped with bowtie2 (Langmead and Salzberg 2012) onto a pre-existing assembly forming so-called "anchored" reads, which are then used to scaffold sequences using a heuristic algorithm (Supplemental Figure 1). Finally, each scaffold is locally assembled using anchored, unmasked reads. The KILAPE pipeline performs read pre-processing, k-mer counting and repeat masking, read mapping to an initial *de novo* assembly, scaffolding and local assembly/gap closure. Starting with the assembly A (Table 7), KILAPE was run three times: first, with 8 kb 454 mate-pairs from fish1 and fish3 followed by two rounds with 20 kb 454 mate-pairs from fish3.

Additionally, gaps were closed with the programs GapFiller (Boetzer and Pirovano 2012) and GapCloser (Luo et al. 2012). For this, three Illumina sequencing datasets with insert sizes of either 170 bp (fish31, run "110404"), 300 bp (fish1, "100512") or 3 kb (fish3, "110118_mp") were mapped with bowtie2 and the parameter "--fast" onto the assembly B (Table 7). All read pairs mapping properly at least 5 kb away from either flank of a scaffold were discarded. The remaining read pairs served as input for gap filling. GapFiller (1.10) was modified so that it also uses bowtie2 for mapping and was run with the three filtered read sets with default parameters except for "-i" (number of iterations), which was set to 6. The resulting improved assembly was passed to GapCloser (1.12) to further close gaps using the same three filtered Illumina read sets and default parameters.

2.1.6 Ordering Scaffolds with Optical Mapping

Whole genome mapping was performed by OpGen producing maptigs which represent a consensus of restriction fragment patterns covering a specific genomic region. These maptigs allow ordering and orienting of sequence scaffolds from the assembly B (Table 7). According to OpGen's requirements in respect to minimal contig length and maximal content of Ns <5%, the scaffolds of the assembly B were split at N-stretches longer than 5 kb, thereby resulting in a set of so-called "OpGen scaffolds". OpGen scaffolds longer than 40 kb were submitted together with the cultured cells to OpGen for optical map creation. There, the extracted high molecular weight DNA was sheared into long molecules, which were stretched and fixed on a glass surface. The restriction enzyme *BamHI* was used to cut the DNA at the specific restriction site GGATCC. This was performed in parallel for several thousand DNA molecules per analysis run. Restriction patterns of each digested DNA molecule were

⁸ <https://github.com/bdownie/KILAPE>

detected by optical methods and saved as “Single Molecule Restriction Maps” (SMRMs) in a proprietary file format including series of fragment lengths and confidence scores. To assemble the SMRMs to maptigs, OpGen scaffolds >250 kb were *in silico* digested to serve as seeds for maptig assembly. This was done in an iterative process where SMRMs are aligned to the seed/maptig and a consensus for the elongated ends is calculated. All maptigs that went at least eight times through this process were kept for scaffold-maptig alignment. For this, OpGen aligned all scaffolds >40 kb to the maptigs using proprietary software and delivered the raw SMRMs, the maptigs and the scaffold-maptig-alignments.

To create optical map-based “superscaffolds”, the original assembly B scaffolds were ordered and oriented according to the aligned OpGen scaffolds (>40 kb) along the maptigs. In this step, also a comparison between optical map-based ordering and sequencing-based scaffolding was done. Inconsistencies were identified and manually resolved either by removing the scaffold-maptig-alignment or by breaking assembly B scaffolds at appropriate gaps.

Finally, long range pairing information of Sanger-sequenced BAC insert ends was utilized to place additional scaffolds in gaps within superscaffolds. For this, BAC end sequences, superscaffolds and the remaining scaffolds were repeat-masked with RepeatMasker (Smit et al. 1996-2015) using an intermediate version of the *N. furzeri* repeat library. The BAC insert ends were aligned to the scaffolds and superscaffolds using BLAST (Altschul et al. 1990). If one genomic insert end of a BAC aligned to a scaffold and the other insert end aligned to a region in a superscaffold next to a gap large enough to contain the scaffold, the scaffold was placed within this gap.

In the assembly of maptigs by OpGen, certain patterns of SMRM arrangements at the ends of maptigs resulted in three types of end annotations: (i) “chromosome end”, (ii) “next to big fragment region” or (iii) “next to potential repetitive region” (see examples in Figure 2). A coverage of SMRMs of at least 30-fold was recommended by OpGen to be sufficient for a reliable prediction of chromosome ends, but lower-coverage annotations were delivered by OpGen as well. The chromosome end annotations were manually inspected and then used as support when ordering superscaffolds in the following assembly steps.

2.1.7 Linkage Map-Based Scaffolding

Three genetic linkage maps were utilized. The first map (Kirschner et al. 2012) was the main resource to build genetic scaffolds (GSCs) while the second (E. Ng’oma, personal communication) and the third map (Ng’oma et al. 2014) were used to complement the first map. All marker sequences were aligned with WU-BLASTn to genome assembly C with parameters “M=1 N=-3 Q=3 R=3 W=30 E=1E-60 V=10 B=10 wordmask=seg lcmask spoutmax=1 hspsepSmax=1000“. To exclude potential misassignments of genetic markers, repeats in the assembly and in the marker sequences were masked with RepeatMasker and an intermediate *N. furzeri* repeat library created

by RepARK. The results were filtered for a minimum sequence identity of 90% and only the best high scoring pair of the best hit per marker sequence was kept. The alignments of the marker sequences to the scaffolds and superscaffolds were visually inspected to identify possible marker misplacements. Superscaffolds and scaffolds were then grouped into genetic scaffolds. Here, the following rules were applied (note that for the sake of convenience the term scaffold is also used for superscaffolds):

- A scaffold is assigned to a linkage group if the large majority of its markers belong to that LG.
- The scaffolds are ordered according to the marker order in the LG.
- A scaffold is reverse-complemented if its marker order is reversed.
- Two LGs are assigned to a single GSC if their markers map to two adjacent parts of the same scaffold.
- Chromosome end annotations provided by optical mapping are used to validate putative scaffold position or orientation.
- All scaffolds that were assigned to a GSC are connected by runs of 100,000 Ns.

For the graphical representation of the marker alignments, the different genetic maps and the sequence ordering within the GSCs, the software package *Circos* (0.67-5) was used (Krzywinski et al. 2009).

2.1.8 Building Synteny Groups

To build synteny groups (SGRs), the genomes of the fish species *Oryzias latipes* (medaka) and *Gasterosteus aculeatus* (stickleback) were analyzed for conserved gene synteny to the *N. furzeri* genome assembly D (Table 7). Medaka and stickleback were chosen because they were at the time the most closely related fish species with a chromosome-level genome assembly available. Stickleback (assembly “BROAD S1”) and medaka (assembly “HdrR”) genome sequences and genes were downloaded from Ensembl⁹. For *N. furzeri*, the gene models had been initially built based on the assembly C (see chapter 2.1.6) but were then converted to assembly D using the UCSC *liftOver* tool (Kuhn et al. 2013). To identify syntenic regions between two species, *mercator* (Dewey 2007) was used. *Mercator* first compares all translated CDS exons of one species versus all of the other species using BLAT (Kent 2002), then builds a graph of exons where significantly aligning exons are connected and finally reconstructs orthologous/syntenic segments. Consequently, regions in assembly D were identified which show conserved synteny and gene order to both medaka and stickleback. Yet unassigned superscaffolds were joined with GSCs or, where necessary, incorporated into GSCs. Such joins were indicated by runs of 100,000 Ns

2.1.9 Additional PacBio-Based Analyses

PBJelly (from PBSuite 15.2.20) (English et al. 2012) was used to fill gaps of the assembly E (Table 7). For this, PBJelly mapped PacBio subreads resulting from WGS sequencing using BLASR (Chaisson and Tesler 2012) onto the assembly E and extended the flanks of gaps towards their centre.

⁹ Database version 74, <http://www.ensembl.org/>

However, the resulting gap fillings were only done as part of assembly evaluations and not incorporated into the final assembly E.

In addition to the gaps filled by PBJelly, sequences of 254 gaps in assembly E (509 kb) were retrieved from BACs independently sequenced and assembled with PacBio data, as described in Reichwald et al. (Reichwald et al. 2015). In brief, PacBio subreads were assembled using the HGAP approach from the SMRT Analysis package and the contigs were then aligned to assembly E. When a BAC contig spanned a gap in assembly E, this region was extracted from that BAC contig.

2.2 The Repeat Identification Pipeline RepARK

2.2.1 The *D. melanogaster* Genome

D. melanogaster has four chromosome pairs of which the first is either a XX or a XY chromosome pair while the other three are the chromosome pairs 2, 3 and 4. The *D. melanogaster* genome assembly (Adams et al. 2000; Celniker et al. 2002) was downloaded from FlyBase¹⁰. Its size is 170 Mb and it contains 15 sequence entries. These are: the left and right arms of chromosomes 2 and 3 (2L, 2R, 3L, 3R), chromosome X (X) plus the corresponding heterochromatin content of these chromosomes (2LHet, 2RHet, 3LHet, 3RHet, XHet), chromosome Y only as heterochromatin (YHet), the mini chromosome 4 (4), the mitochondrial genome (M), plus two additional pseudo-chromosomes (U, Uextra).

2.2.2 Available Repeat Libraries for *D. melanogaster*

Two resources of repeat libraries were used. First, from RepBase Update (release 20120418), 412 repeat consensi were fetched with the RepeatMasker utility `queryRepeatDatabase.pl` (“-species “*drosophila melanogaster*””) which extracts *D. melanogaster*-specific repeats (26) as well as ancestral repeat consensi that are shared with species of the same clade. After removing repeats marked as “low-complexity”, 249 repeats remained and built the library “DmRepBase”. A second repeat library created in the Drosophila 12 Genomes Project (Clark et al. 2007) using ReAS was also downloaded¹¹. This library contains 391 repeat consensi without any classification (“ReASLib”). Statistics of both libraries are given in Table 4.

¹⁰ version R5.43, ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.43_FB2012_01/fasta/dmel-all-chromosome-r5.43.fasta.gz

¹¹ ftp://ftp.genomics.org.cn/pub/ReAS/drosophila/v2/consensus_fasta/dmel.con.fa.gz

Table 4: Available *D. melanogaster* Repeat Libraries.

	DmRepBase	ReASLib
Number of consensi	249	391
Total length [Mb]	0.7	0.96
Min/max length [bp]	52/14,477	101/12,876
N50 length [bp]	5,402	4,757
N90 length [bp]	1,750	1,247

2.2.3 Sequencing Data

For constructing the repeat libraries with RepARK (“Repetitive motif detection by Assembly of Repetitive K-mers”), three short-read datasets with an approximated genome coverage of 40-fold were used for RepARK evaluation; two for *D. melanogaster* and one for human. A “simulated” *D. melanogaster* genomic sequence dataset was created with MAQ¹² (0.7.1) without mutations or indels using the entire *D. melanogaster* genome assembly release R5.43 as input. MAQ was trained with a typical 101 bp Illumina dataset of the *N. furzeri* project. The library statistics for this as well as for the following datasets are given in Table 5. The experimentally obtained Illumina read dataset of *D. melanogaster* was retrieved from the Short Read Archive (SRA¹³). It consists of two sequencing runs of a fly of the stock¹⁴, which was used for the release 5 genome assembly (Langley 2011). Both *D. melanogaster* datasets were error-corrected with QUAKE (0.3.4) (Kelley et al. 2010), using default settings and “k=17”. The human Illumina reads derived from sequencing of a lymphoblastoid cell line (Coriell Institute, GM12878) were also downloaded from the SRA and used directly without error correction.

Table 5: Sequence Data Used for the RepARK Development and Evaluation.

	<i>D. melanogaster</i> , simulated	<i>D. melanogaster</i> , real	Human
Number of reads [million]	68	83	1,307
Total size [Gb]	6.9	6.8	132
Read length [bp]	101	82	101
Insert size [bp]	400 and 2,500	350	180
Sequence source or accession numbers	<i>D. melanogaster</i> genome assembly R5.43	SRX040484 (ycnbwsp_2), SRX040486 (ycnbwsp_7-HE)	SRR067780, SRR067784, SRR067785, SRR067787, SRR067789, SRR067791, SRR067792, SRR067793

2.2.4 Building the RepARK Repeat Libraries

For the *de novo* repeat library creation based on NGS data, k-mers of the three datasets were first counted with Jellyfish (1.1.6) (Marcais and Kingsford 2011) using a k-mer size of 31 (“-m 31 --both-strands”). The threshold above which a k-mer is considered repeat-derived (i.e.

¹² <http://maq.sourceforge.net>

¹³ <http://www.ncbi.nlm.nih.gov/sra>

¹⁴ <http://flybase.org/reports/FBst0002057.html>

occurring more than once genome-wide) was individually predicted for each dataset. For this, a histogram of k-mer frequencies is calculated. A typical k-mer histogram can be divided into three parts. At the left, all k-mers are listed that contain errors and thus occur at a low frequency. The central part resembles a Poisson-like distribution of the major fraction of k-mers around a peak which is usually at or near the expected sequencing coverage. To the right of this peak are all k-mers that occur many-fold across the dataset. The latter fraction is used for the repeat library assembly. To determine a defined threshold for a separation of k-mers, a linear function is fit to the slope of the descending segment of the unique k-mer fraction (middle part). K-mers with a frequency above which the projected linear function crosses the x-axis are expected to occur more than once genome wide. To further avoid possible contaminations of the abundant k-mer set by unique sequences, this value is doubled, and k-mers with a frequency above this threshold are considered abundant. Abundant k-mers were isolated and independently assembled using the *de novo* DBG assembly programs CLC Assembly Cell (abbreviated CLC, 4.0) or Velvet (1.2.08) both with the k-mer parameter set to 29, resulting in four RepARK *de novo* repeat libraries of *D. melanogaster* and one library for human.

2.2.5 Building Additional Repeat Libraries for Comparison

Additionally, repeat libraries for both real and simulated *D. melanogaster* datasets were *de novo* generated using two established methods. First, to apply the classical approach of repeat library creation, a *de novo* genome assembly was performed with Velvet using either the simulated or the real *D. melanogaster* reads (Assembly statistics are given in Supplemental Table 1). Then RepeatScout was applied to predict repetitive consensi in these genome assemblies. Second, wgs-assembler (Myers 2000) was used to assemble the same read sets and thereby generate surrogates representing those contigs determined to be repetitive by the program.

2.2.6 Completeness Calculation

The completeness value (or fraction of representation) reflects to which fraction a given gold-standard RepBase repeat is represented in a *de novo* repeat library. Because repeats in the *de novo* libraries are fragmented, one RepBase repeat is represented by several *de novo* consensi. To calculate completeness the *de novo* library was analyzed by RepeatMasker and its 249 *D. melanogaster* repeats. The number of masked nucleotides was calculated for each RepBase repeat (every masked nucleotide was counted only once) and divided by the full length of this RepBase repeat.

2.2.7 Repeat Classification with TEclass

The repeat consensi were classified with TEclass (2.1) (Abrusan et al. 2009) using the default training set based on all available RepBase (release 15.07) repeats. Because TEclass compares patterns of oligomer (k-mer) frequencies of this training set with those of the repeat consensi, a minimum length of 50 bp is mandatory to predict the repeat class with a certain confidence. Due to this restriction, repeat consensi below 50 bp were discarded from the repeat libraries prior to TEclass analysis.

2.2.8 Mapping and Repeat Masking

D. melanogaster or human repeat libraries were mapped to their respective reference genome using BLAT (.34) with default options including “-extendThroughN” to align over stretches of Ns in the reference and “-minIdentity=50” to retain lower identity hits in the raw results. The resulting psl files were further filtered for a higher minimum identity as mentioned in the results section. Repeat masking was performed with RepeatMasker (4.0.0) with the default parameters and either *D. melanogaster* or human repeats from RepBase (DmRepBase, release 20120418) or the specified *de novo* repeat library.

2.2.9 Retrieving Known SDs and Comparison to the *de novo* Repeat Consensi

The coordinates of SDs identified in release 5 of the *D. melanogaster* genome assembly were downloaded¹⁵. Based on the coordinates, respective SD sequences were extracted from the reference genome and then masked with DmRepBase to exclude “normal repeats” biasing this analysis. After masking, 3.09 Mb SD regions without RepBase-repeats remained. Additionally, each repeat library was also masked separately with DmRepBase. To determine the SD fraction that each library can identify, RepeatMasker was employed again, this time to mask the remaining (unmasked) SD fraction with each masked repeat library analyzed in this study.

2.3 Identification and Classification of Genomic *N. furzeri* Repeats

2.3.1 Genome Assembly-Based Library Creation

To build the genome assembly-based library, RepeatModeler (1.0.7) was run on the *N. furzeri* assembly C (Table 7) with default parameters.

2.3.2 Read Data-Based Library Creation

The construction of the Sanger-based repeat library was previously described in detail (Koch 2010). In brief, fragments of a WGS library of *N. furzeri* were sequenced with Sanger technology from both ends and overlapping end sequences were assembled into contigs (Table 2). These sequences were scanned for TRs by TRF and subsequently analyzed with RepeatMasker to find homologies to reference repeats in RepBase. After masking the TRs and the RepBase repeats with Ns, the *de novo* identification program RepeatScout was used to identify *N. furzeri*-specific repeats.

Of the sequencing runs 110404 and 110421 from female specimen fish31 the forward reads of the down-sampled datasets were extracted (48 Gb, 25-fold coverage depth, Table 6) and 31-mers were counted with jellyfish (1.1.6). Of the resulting 31-mer distribution histogram, RepARK calculated a threshold designating 31-mers occurring at least 66 times as repeat derived. These abundant 31-mers were isolated and assembled using CLC (4.0) with a k-mer size of 29 (“-w 29”) to build the basic RepARKLib for *N. furzeri*.

¹⁵ <http://humanparalogy.gs.washington.edu/dm3/dm3wgac.html>

2.3.3 Size Selection and Redundancy Reduction of Repeat Libraries

All sequences shorter than 80 bp were excluded from the repeat libraries, according to the 80-80-80 rule (Wicker et al. 2007). To remove redundant sequences within different repeat libraries, CD-HIT-EST (4.5.4) (Li and Godzik 2006) was used with parameters “-c .80” (sequence identity threshold) and “-n 8” (word length) to cluster the sequences. For each cluster, CD-HIT-EST determines a master sequence that represents the similar sequences of that cluster. These master sequences were regarded repeat consensi and used for further analysis.

2.3.4 Repeat Classification

For the first iteration of the classification of repeat consensi, Censor (version “censor.ncbi V4.2.28”) was applied to find similarities to reference repeats in RepBase (release 17.04). Consensi that remained unclassified were re-analyzed with Censor using 2,881 classified repeats from FishRepLib, which is a teleost-specific library containing repeat consensi from stickleback, zebrafish, Nile tilapia, ghostshark, and platyfish (Chalopin et al. 2015). The actual assignment of a class to a repeat consensus was carried out with the Perl script `annotateFromCensorMap.pl` (D. Chalopin, personal communication) using the “map table” output of Censor. Only repeat consensi that were covered by a classified reference repeat over at least 50% of their length were classified this way. After this, the subsequent classification efforts were performed manually. RNAfold and RNAPlot from the Vienna RNA Package (1.6.1) (Hofacker et al. 1994) were used to predict the potential secondary structure for each repeat sequence with a length ≥ 250 bp. These structures were then inspected for tRNA motifs or hairpin like structures of which the latter are characteristic for MITE repeats. Additionally, consensus sequences of two *N. furzeri*-specific highly abundant tandem repeats (Reichwald et al. 2009) were aligned to the remaining unknown sequences with BLASTn. Hits were filtered for 80% identity and 80% length coverage of the TR consensus sequence and the respective repeat consensus was marked as “#Satellite”. A tBLASTn alignment against six helicase protein sequences finally identified Helitrons.

2.3.5 Genome Assembly Repeat Annotation

For comparing the different repeat libraries and for annotating repeats genome-wide, RepeatMasker (4.0.5) together with the search engine “ncbi/RMblast” (parameter “-e ncbi”) was applied to the genome assembly using the different repeat libraries. The repetitive fraction was calculated based on both the assembly with (1.24 Gb) and without (0.86 Gb) regions of Ns.

For the final repeat annotation, the genome assembly was first analyzed for TRs using TRF with parameters “2 7 7 80 10 50 2000”. TR occurrences in the resulting .dat file were categorized into microsatellites (1-5 bp), minisatellites (6-99 bp) and satellites (≥ 100 bp). As TRF sometimes reports shorter TR motifs within longer ones, merging of the intervals was performed with `bedtools merge` (Quinlan and Hall 2010) for each of the three categories thereby removing redundancy. In the second step, RepeatMasker together with the final *N. furzeri* repeat library

(“CombinedLib”) masked the TR-masked genome assembly. To calculate the different DR classes detected in this step, the script `buildSummary.pl` (provided with RepeatMasker) was applied to the primary masking output file and the fractions of the major classes LINE, DNA, LTR, SINE as well as of other and unknown repeats were extracted.

The evolutionary history of TEs was calculated based on the previous RepeatMasker analysis. Here, it is assumed that a consensus from a repeat library is an approximation of the ancestral repeat element and the individual repeat copies in the genome changed/diverged during molecular evolution. This difference is represented by the Kimura distance K (Kimura 1980) which is calculated for each consensus based on its alignment to the reference with the equation $K = -1/2 \ln(1 - 2p - q) - 1/4 \ln(1 - 2q)$ (where p is the proportion of sites with transitions and q is the proportion of sites with transversions). Elements with a lower distance show greater similarity between the copies and the consensus and are therefore considered younger compared to those with higher distances. A younger repeat is assumed to be recently incorporated, still active and proliferative in the genome compared to an older one whose activity can date back millions of years. Events with a high activity of a certain repeat class are called bursts of transposition. For this analysis, the Perl script `RepeatLandscape.pl`¹⁶ was applied to the alignment output file. This script was modified to cope with different spellings of the same superfamily. It extracts Kimura distances from the alignments file and sums up the genomic fraction of each superfamily over 50 categories ranging from Kimura distance 0 to 50. From the resulting dataset, all superfamilies were discarded that show a genomic fraction of less than 0.2%. Genomic fraction values from this analysis are usually slightly higher than those obtained by `buildSummary.pl` because `RepeatLandscape.pl` also counts overlaps of elements twice.

2.3.6 Repeat Analysis of PacBio-Filled Gaps and Completeness Estimation of the Genome Assembly

Sequences of gaps filled by either PacBio data were categorized into bins with increasing gap length (1-99; 100-299; 300-899; 900-2,699; 2,700-8,099; 8,100-24,299; 24,300-72,899 bp). Each bin was first analyzed with TRF to detect TRs. Second, RepeatMasker together with the CombinedLib was applied to the TR-masked gap sequences. A logarithmic function was fitted to the resulting repeat content, which was then used to estimate the repeat content of longer gaps.

¹⁶ <https://github.com/caballero/RepeatLandscape>

The fraction of unique bases in the genome assembly was calculated as

$$UniqueFrac = \frac{U_A}{U_G} \cdot 100$$

with

$$U_A = A - N_A - R_A$$

being the unique bases in the assembly and

$$U_G = U_g + U_A + U_{!A}$$

the unique bases in the genome. A is the assembly size, N_A the Ns in the assembly, R_A the repeat bases in the assembly, U_g the approximated number of unique bases in the assembly gaps, U_A the number of unique bases in the assembly and $U_{!A}$ the number of unique bases estimate in the not assembled genomic sequence. The latter can be calculated from

$$U_{!A} = G - A - R_{!A}$$

with G being the genome size, A the assembly size and

$$R_{!A} = (G - A) \cdot X_g$$

the number of repeat bases estimate in the not assembled genomic sequence with X_g as the approximated repeat fraction of large gaps. The Fraction of repeat bases in the genome was estimated accordingly with

$$RepeatFrac = \frac{R_A}{R_G} \cdot 100$$

with R_A the repeat bases in the assembly and

$$R_G = R_g + R_A + R_{!A}$$

the repeat bases in the genome with R_g being the approximated number of repeat bases in the assembly gaps.

3 Results

3.1 Assembling a Reference Sequence of the *N. furzeri* Genome

The goal of the *N. furzeri* genome project at the FLI is to provide a high-quality reference sequence of the genome, which will allow making full use of *N. furzeri* as a model to study the genetics and biological determinants of aging and longevity. In this thesis, I dedicated my bioinformatics work on the *de novo* assembly of the genome and a comprehensive analysis of its repeat content.

3.1.1 Sequencing and Data Preparation

For the basic assembly, two paired-end and ten mate-pair libraries were sequenced (Table 3). Because the mate-pair libraries contained up to 92% of duplicons, I removed highly abundant duplicons thus reducing mate-pair coverage from 77.2-fold to 49.9-fold (coverage calculation based on genome size estimate of 1.9 Gb). This filtered mate-pair dataset was used for the ALLPATHS-LG (Gnerre et al. 2011) assembly. The paired-end coverage was down-sampled accordingly from 83-fold to 50-fold to match the program's requirements (Table 6).

Table 6: Sequence Data Used for the ALLPATHS-LG Initial Assembly after Down-Sampling.

Date ID	Accession	Specimen	Reads	Bases	Coverage ^a
Paired-end libraries (insert size 170 bp)					
110404	ERR583470	fish31	423,949,984	42,394,998,400	22.3
110421	ERR583471	fish31	526,050,015	52,605,001,500	27.7
Paired-end total			949,999,999	94,999,999,900	50.0
Mate-pair libraries (insert size ~3 kb)					
110506	ERR583472	fish55	87,771,150	8,012,970,025	4.2
110627	ERR583473	fish55	31,361,100	2,852,102,822	1.5
110627	ERR583474	fish55	280,373,652	25,685,056,349	13.5
110627	ERR583475	fish55	76,942,520	7,132,461,895	3.8
110627	ERR583476	fish55	36,521,642	3,404,991,464	1.8
110916	ERR583477	fish55	49,571,268	4,714,119,862	2.5
110916	ERR583478	fish55	78,896,064	7,496,651,873	3.9
110916	ERR583479	fish55	97,132,382	9,266,977,586	4.9
111103	ERR583480	fish55	150,617,238	14,202,329,488	7.5
111103	ERR583481	fish55	129,973,960	12,051,495,984	6.3
Mate-pair total			1,019,160,976	94,819,157,348	49.9

^a Coverage calculations are based on 1.9 Gb genome size.

3.1.2 Genome Assembly

As the *N. furzeri* genome is complex and repeat-rich, a multi-step strategy was applied. It included an (A) initial *de novo* assembly, which was step-wise improved by (B) scaffolding, (C) optical mapping, (D) integration of genetic linkage maps and (E) comparative synteny mapping (Table 7). I performed steps A, C and D while B and E were done by colleagues (Reichwald et al. 2015) and therefore I will give only a short summary of the latter.

Table 7: Assembly Statistics.

Assembly Step		Number of sequences	number of Ns	Total length [bp]	Shortest / longest sequence [bp]	N50 length ^a [bp]
A	ALLPATHS-LG (contigs)	126,539	88,013 (0.01%)	811,928,617	46 / 104,905	10,545
	ALLPATHS-LG (scaffolds)	15,930	88,983,326 (9.88%)	900,823,930	886 / 1,451,049	132,538
B	KILAPE + gap filling (contigs) ^b	10,894	62,039,618 (6.75%)	918,829,636	886 / 2,200,336	259,182
	KILAPE + gap filling (scaffolds)	7,675	86,805,836 (9.20%)	943,595,854	886 / 3,869,209	494,454
C	Optical map integration	6,012	374,109,457 (30.39%)	1,230,898,532	886 / 44,272,285	15,858,201
D	Genetic map integration	5,924	382,909,457 (30.89%)	1,239,698,532	886 / 96,068,516	48,234,189
E	Comparative synteny mapping	5,896	385,709,457 ^c (31.04%)	1,242,498,532 ^c	886 / 98,476,147	57,367,160

^a The sequence size above which half the total assembly size is represented. ^b Scaffolds of the assembly B were split at gaps longer than 5 kb to obtain contigs statistics. ^c This number is used for assembly completeness calculation in chapter 3.3.6. [modified from Table 1 (Reichwald et al. 2015)]

3.1.2.1 Basic Assembly with ALLPATHS-LG - Assembly A

In a very first test, ALLPATHS-LG (version 38405) was run at an early time point when all paired-end data, but only 32-fold mate-pair coverage was available. This assembly resulted in 22,349 scaffolds with a total length of 861 Mb and an N50 length of 87 kb.

The actual *N. furzeri* genome *de novo* assembly A was built by using ALLPATHS-LG (version 42316) when the required coverage of the libraries was available and the read data was prepared as described in 2.1.3. The runtime of the program was 271 h with a peak memory usage of 422 GB. Many of the program's steps were distributed between the available 64 CPUs resulting in an effective parallelization factor of 30.7. Along with the genome assembly, ALLPATHS-LG estimated a genome size of 1.3 Gb and a repeat content of 53% based on a 25-mer distribution analysis of the paired-end reads. Overall, 126,539 contigs were assembled and connected to 15,930 scaffolds amounting to 900 Mb. The N50 length which is the length of a scaffold above which half of the total assembly size is represented was 132 kb and the maximum length was 1.5 Mb. Due to scaffolding, assembly A contains 89 M ambiguous positions (10%) that represent mostly gaps between contigs (Table 7).

3.1.2.2 Further Scaffolding with KILAPE and Gap Filling - Assembly B

For further scaffolding, the 8 kb and 20 kb mate-pairs from 454 sequencing were integrated into the assembly A using the KILAPE pipeline. First, KILAPE predicted three classes of k-mers based on a spectrum of 17-mer counts: (1) "erroneous" k-mers occurring with a frequency of 1 to 3, (2) "unique" k-mers with a frequency of 4 to 193 and (3) "repetitive" k-mers with a frequency equal or above 194. KILAPE then masked the erroneous and repetitive k-mers in the paired 454 reads and subsequently scaffolds the input assembly. To reduce the number of Ns introduced upon scaffolding with ALLPATHS-LG and KILAPE, the programs GapFiller (Boetzer and Pirovano 2012) and GapCloser (Luo et al. 2012) were applied using the Illumina datasets "110404" (Table 3), "100512"

and “110118_mp” (Table 2). This reduced the number of Ns in the post-KILAPE assembly from 134 Mb to 86.8 Mb (9.2% of assembly B; data of this intermediate step not shown). After this step, the number of scaffolds was reduced by half, with a total assembly length of 944 Mb. The N50 length increased to 494 kb and the longest scaffold has a size of 3.9 Mb (Table 7). However, the number of novel unambiguous bases increased by 40 Mb, mainly due to extensive long-range scaffolding.

3.1.2.3 Superscaffolding with Optical Maps - Assembly C

The *de novo* optical map of the *N. furzeri* genome was generated by OpGen. From the stretched and digested genomic DNA, OpGen performed 38 data collection runs to obtain 2.7 M SMRMs with an average molecule size of 286.7 kb and an average *Bam*HI fragment size of 14.6 kb. According to OpGen’s requirements, the scaffolds of assembly B were split at gaps >5 kb (Table 8), here called “OpGen scaffolds”. A subset of 1,063 OpGen scaffolds >250 kb (Table 8) was used by OpGen as seeds for creating 106 maptigs with a SMRM coverage depth of at least 30-fold, an average length of 9.2 Mb and a total length of 975 Mb (Table 9). A set of 4,519 OpGen scaffolds >40 kb (Table 8) was aligned by OpGen to the maptigs and 2,677 (59%) were placed onto the 106 maptigs (exemplified in Figure 1). Based on these placements, 105 so-called superscaffolds with an N50 length of 16.6 Mb were built, which represent the major fraction of assembly C (1.07 Gb, 86.7%; Table 7). There was one maptig (3.3 Mb), onto which only a single OpGen scaffold was placed and therefore no superscaffold was created. Accordingly, 5,964 scaffolds of the assembly B remained unplaced. Of those, 57 scaffolds were placed in gaps within superscaffolds by mapping of genomic insert ends of BACs.

Table 8: OpGen Scaffolds of Assembly B, Broken at Gaps >5 kb and Used for Optical Mapping.

OpGen scaffolds	All	>40 kb	>250 kb
Number	10,894	4,519	1,063
Total length [Mb]	919	847	471
Number of Ns [Mb] (percentage)	62 (6.75%)	56 (6.56%)	27 (5.71%)
Shortest / longest [kb]	0.9 / 2,200	40 / 2,200	250 / 2,200
Average length [kb]	84	187	443
N50 length [kb]	259	282	451

Table 9: Maptigs Obtained by SMRM *de novo* Assembly.

Maptigs	Assembly
Number	106
Total length [Mb]	975.3
Average size [Mb]	9.2
Largest [Mb]	38.8
N50 length [Mb] ^a	5.9
Potential chromosome ends	24

^a N50 length was calculated based on 1.6 Gb genome size.

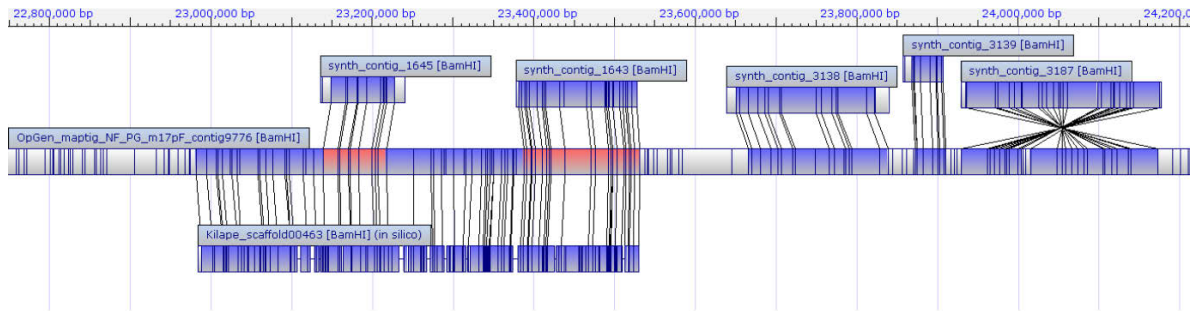


Figure 1: Alignment of Optical Mapping and Sequencing Data for Part of *N. furzeri* superscaffold00010.

Horizontal bars with vertical lines represent either a maptig (middle) or *in-silico* digested OpGen scaffolds (“synth_contig_XXXX”, top) or assembly scaffolds (“KILAPE_scaffoldXXXXX”, bottom). The vertical lines within restriction maps are *Bam*HI restriction sites while those between maps represent alignments of restriction fragments. Blue fragments depict single alignments while red is for multiple alignments. Short horizontal lines in the middle of the assembly scaffold represent stretches of Ns within the sequence. The OpGen scaffolds “1643” and “1645” were originally derived from assembly scaffold “00463”. The scaffold “00463” was added to the alignment to illustrate that it belongs to this location within the later superscaffold00010. Assembly scaffolds corresponding to OpGen scaffolds “3138”, “3139” and “3187”, are not shown here. This screenshot was taken from OpGen’s MapSolver software.

By building superscaffolds, the fraction of Ns in the assembly rose from 9.2% to 30.4%, because no additional sequence information was added; instead, only scaffolds were ordered according to the topology of maptigs and respective gaps were filled by Ns. The resulting assembly C has a size of 1.23 Gb, an N50 length of 15.9 Mb and the longest superscaffold is 44.3 Mb (Table 7).

In the process of creating maptigs from SMRMs, 186 maptig ends were annotated and categorized by OpGen into three types: (i) 24 ends are putative “chromosome ends” which is reflected in a clear coverage drop of SMRMs (Figure 2A). (ii) 158 ends are a “next to big fragment region” where the SMRMs do not have restriction sites within a relatively long region (Figure 2B). (iii) Four ends are a “next to potential repetitive region” which is reflected by many short successive fragments at the end of SMRMs (Figure 2C). For each chromosome end annotation, the SMRM coverage was provided by OpGen, and ranges from 10-fold to 90-fold serving as an indicator for the reliability of this annotation. The annotation “chromosome end” was integrated into the following assembly steps whereas the other two were not used.

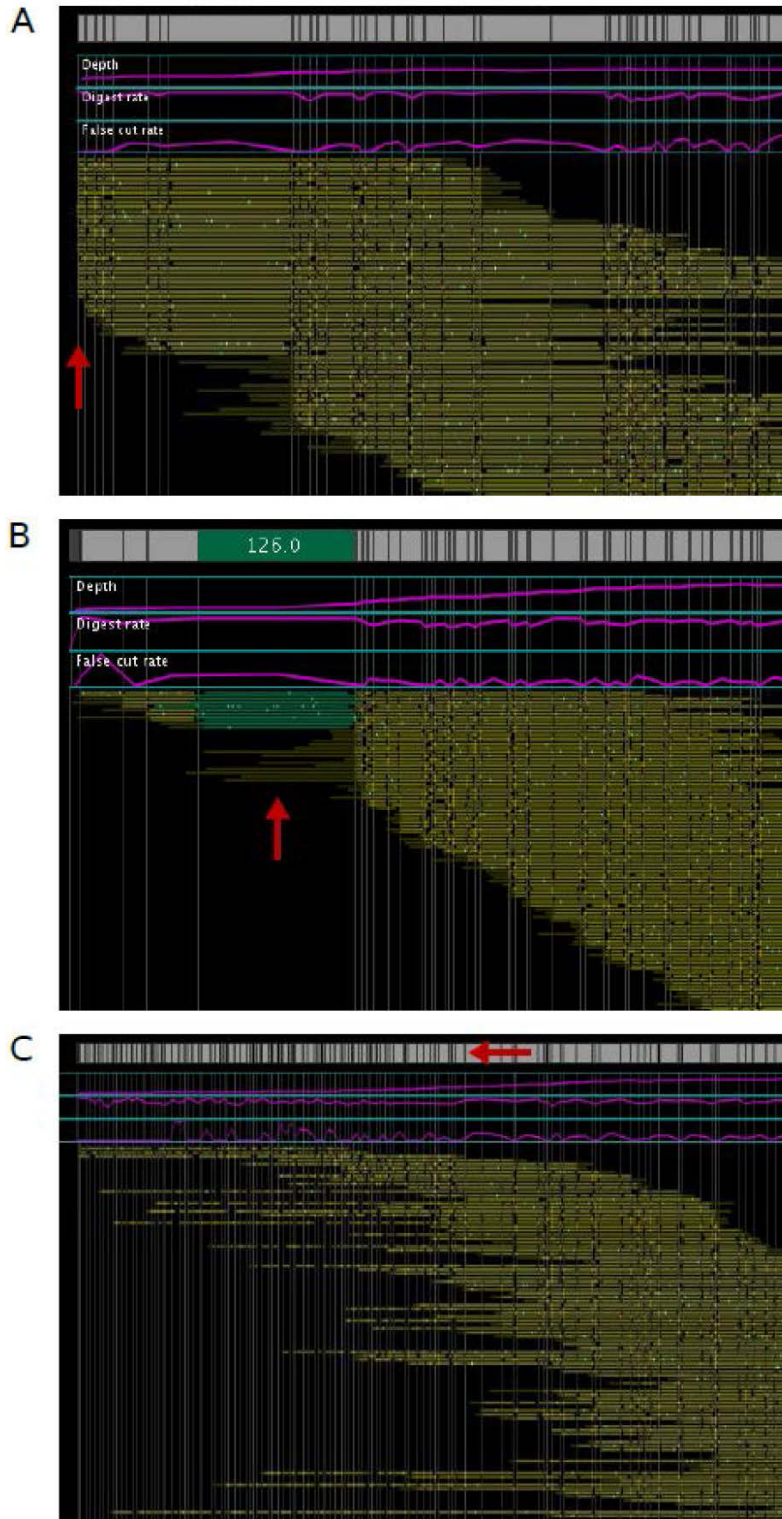


Figure 2: Annotation of Maptig Ends by OpGen.

Three examples of maptig end annotation are shown. (A) The red arrow marks a genomic region where there is a clear drop in restriction fragment coverage (i.e., from >30-fold to none); this represents a telomeric or sub-telomeric region and is annotated as “chromosome end”. (B) The red arrow marks a long region where no restriction took place; therefore this is annotated as “next to big fragment region”. (C) The red arrow marks the beginning and points into the direction of a long sequence of very short restriction fragments, which are assumed to be derived from repeats. In the top row of each plot (gray boxes), a maptig is shown which represents the consensus of BamHI restriction fragments that are depicted below as overlapping green bars. In those, small light green boxes indicate false cuts and yellow boxes show uncut restriction sites. Three purple lines below the maptig show (from top to bottom): restriction fragment coverage (Depth), digestion rate (Digestion) and false cut rate.

3.1.2.4 Linkage Map-Based Scaffolding - Assembly D

To further order scaffolds and superscaffolds, genetic linkage information of three different linkage maps was integrated. The linkage map created in 2012 by Kirschner et al. served as the main resource for building GSCs and was originally published with 124 microsatellite and 231 gene-associated markers grouped into 22 LGs (Kirschner et al. 2012). It was later extended by 37 additional unpublished markers resulting in a total of 392 markers. This map will be referred to as “G1” in the

following. Two other linkage maps were used to include additional superscaffolds into the GSCs: one is comprised of 82 microsatellite markers (“G2”, unpublished, Ng’oma) and another contains 233 gene associated markers (“G3”) (Ng’oma et al. 2014). Of the 392 markers in map G1, 387 (99%) showed valid alignments (WU-BLASTn, $p < 1e^{-60}$) to the assembly C (Supplemental Table 2). The alignments were manually inspected and the respective scaffolds/superscaffolds connected to form 19 GSC. Further, relationships between genetic markers, scaffolds, superscaffolds and GSCs were plotted using Circos aiming to facilitate a manual evaluation and to give a clear visual overview. For example, gsc02 was built from six superscaffolds and one scaffold based on unambiguous alignments of all 30 markers of LG2 of the map G1 (abbreviated as “G1_LG2”, top left, Figure 3).

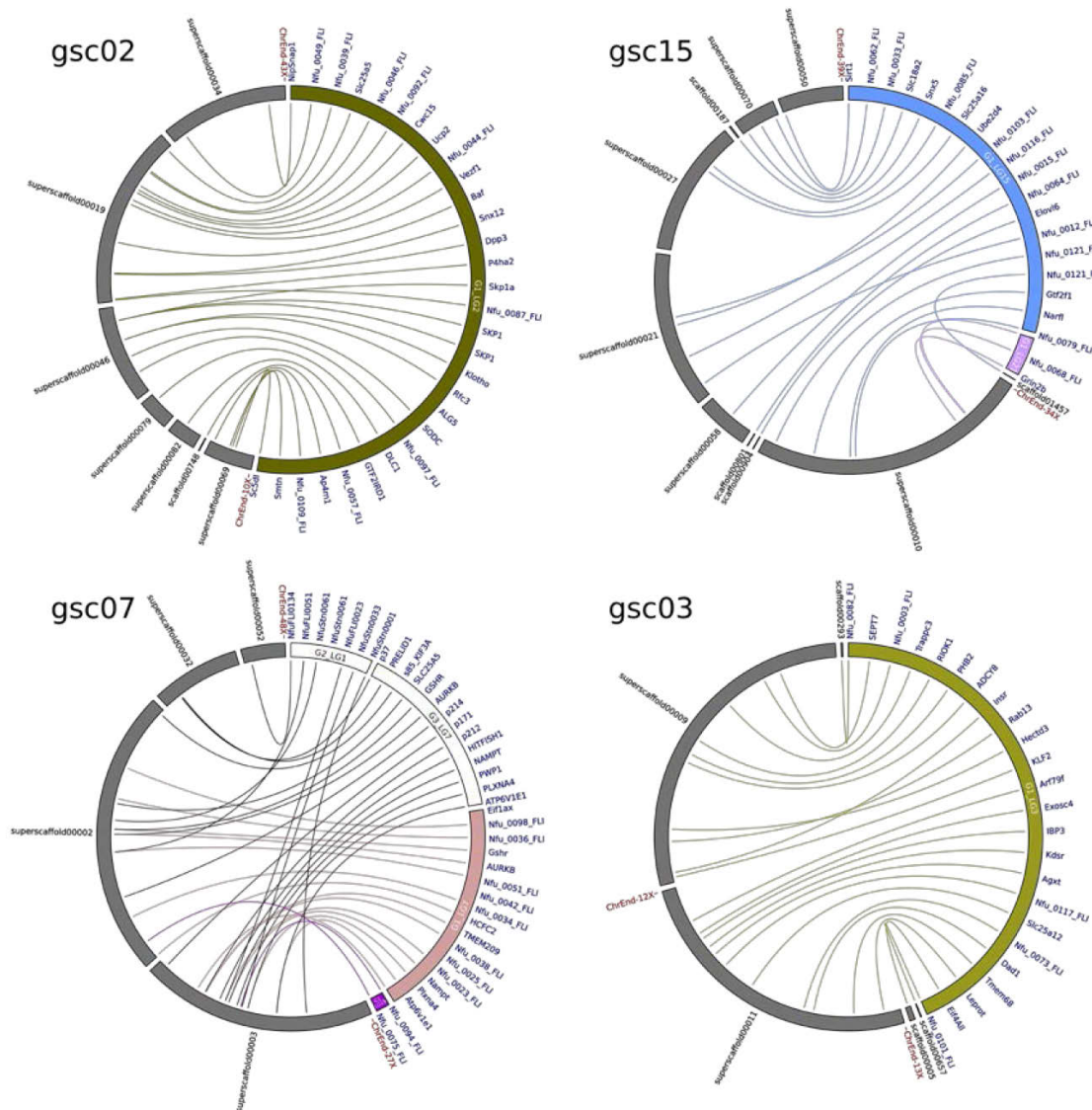


Figure 3: Construction of Genetic Scaffolds.

Four examples of GSCs are shown, in which sequence scaffolds and genetic information are integrated. Scaffolds and superscaffolds from assembly C are depicted in dark gray at the left side of each circle and represent their final order within gsc02, gsc03, gsc07, and gsc15, respectively. LGs with their respective markers are given at the right side of the circle. LGs filled with different colors represent the linkage map G1 while the uncolored groups are from G2 and G3. LGs are named according to their original genetic map; e.g. “G1_LG2” means LG2 of map G1. Lines within the circles represent alignments of markers to scaffolds and superscaffolds. Chromosome end annotations are labeled in red. For convenience, markers of the genetic maps are depicted as evenly distributed and do not reflect their real distances. High-resolution figures of all 19 GSCs are given in Supplemental Figure 5.

Although *N. furzeri* was shown to have 19 chromosomes (Reichwald et al. 2009), all available linkage maps have either 20 or 22 LGs. This suggests that excess LGs will collapse with other LGs to represent a certain chromosome. Accordingly, three LGs of map G1 were joined with other LGs: (i) LG15 and LG21: four markers of G1_LG15 and three of G1_LG21 align to superscaffold00010; these two LGs were merged to gsc15 (top right, Figure 3); (ii) LG19 and LG7: one marker of G1_LG19 aligns to superscaffold00002 and another marker of G1_LG19 to superscaffold00003. Because both superscaffolds have each seven G1_LG7 markers aligned, G1_LG19 was considered to be a fragment of G1_LG7 and therefore both LGs were merged to gsc07 (bottom left, Figure 3); (iii) LG8 and LG22: one marker of G1_LG8 and two markers of G1_LG22 align to superscaffold00017 resulting in gsc08 (Supplemental Figure 5).

To integrate additional superscaffolds and scaffolds into the intermediate G1-based version of the assembly, concordance to linkage maps G2 and G3 was evaluated. In general, there was good agreement between the maps allowing the assignment of six additional superscaffolds to either gsc01 (1), gsc06 (1), gsc07 (2) or to gsc11 (2) (bottom left, Figure 3 and Supplemental Figure 5).

Furthermore, chromosome end annotations provided by OpGen were used to assist in the assembly of the GSCs. In total, 24 chromosome end annotations on 23 superscaffolds were provided, of which 17 were integrated into GSCs. In all cases but one, chromosome end annotations of the superscaffolds coincided with the ends of genetic scaffolds. The only exception was found on gsc03 where superscaffold00011 had two chromosome ends annotated but, based on genetic linkage information, had to be joined to superscaffold00009 (bottom right, Figure 3). Because the marker order of G1_LG3 is consistent in the two superscaffolds and independent FISH analyses support this order, at least one of the two end annotations of superscaffold00011 had to be subject to critical scrutiny. According to OpGen, a chromosome end annotation can be considered reliable when its SMRM coverage depth is 30-fold or larger (which is true for 14 of the 24 chromosome ends). The chromosome ends on superscaffold00011 have a coverage depth of only 12-fold and 13-fold, respectively. Additionally, the signature of the SMRM alignment at the 12-fold end is more similar to a “next to big fragment region” than to that of chromosome ends (Supplemental Figure 2). As these suggested a false-positive annotation with respect to the 12-fold end, it was excluded from analysis.

In total, 74 superscaffolds and 33 scaffolds were assembled into 19 GSCs, which together have a size of 950.8 Mb (76.7% of the entire assembly). Together with the remaining superscaffolds and scaffolds (23.3%), assembly D comprises 1.24 Gb with an N50 length of 48.2 Mb (Table 7).

3.1.2.5 Synteny-Based Scaffolding - Assembly E

The final scaffolding step was aimed at further integrating remaining superscaffolds and scaffolds into the genetic scaffolds to form SGRs. This was accomplished by performing synteny analyses in the genomes of stickleback and medaka. This analysis identified 2,971 regions (726 Mb) in *N. furzeri*, which showed conserved synteny and conserved gene order (i.e. for at least two genes) to medaka and

2,979 *N. furzeri* regions (750 Mb) which showed conserved synteny and order to stickleback. Based on these two synteny maps, *N. furzeri* GSCs were assigned to respective medaka or stickleback chromosomes. Accordingly, 28 superscaffolds were added to the set of GSCs and thus formed 19 SGRs (Table 10). In ten cases, a GSC was split at a certain position to allow inserting a superscaffold at this position (marked with an asterisk in Table 10). For six of these additional superscaffolds, their chromosomal end annotation by OpGen coincided with a SGR end, which increased the number of SGR ends annotated as chromosomal ends to 22 of 38 (58%). The SGRs were named according to their final length, with sgr01 being longest and sgr19 shortest. The 19 SGRs alone (Figure 4) comprise 1,079 Mb and account for 87% of the *N. furzeri* genome assembly E (1.24 Gb with an N50 length of 57.4 Mb, Table 7).

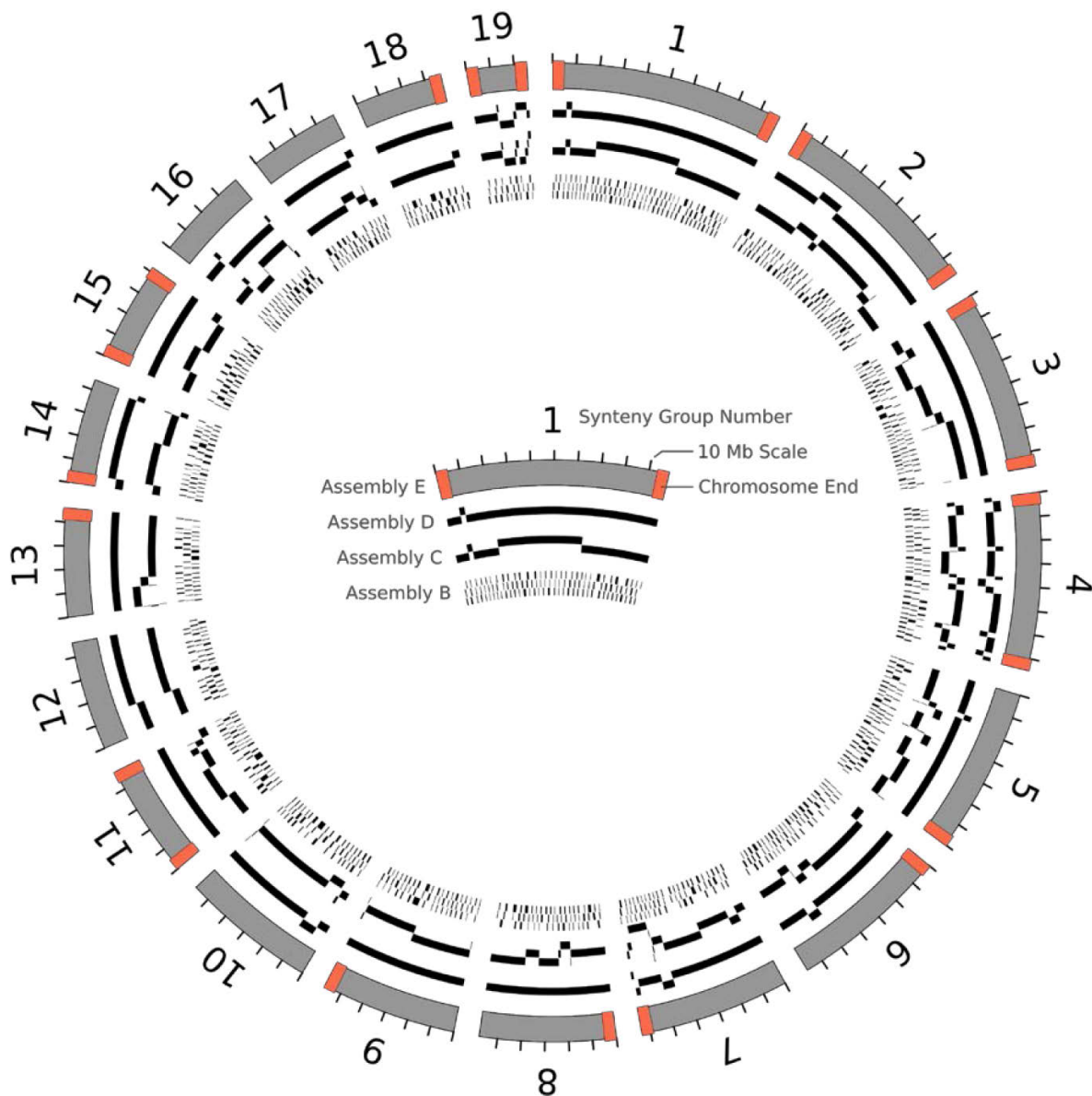


Figure 4: Assembly Overview Including 19 Synteny Groups.

The circles represent the assembly steps of the *N. furzeri* genome; the inner circle represent early and intermediate steps while the outer circle represents the currently most complete assembly. Sequences that could not be anchored to SGRs are not shown. [modified from Figure 1 (Reichwald et al. 2015)]

Table 10: Construction Scheme of the SGRs.

Sequence to assign		Target SGR		
Name	+/- ^a	Name	Start ^b	End ^b
gsc07 ^{End}	+	sgr01	0	6.4
superscaffold00083*	-		6.5	8.8
gsc07 ^{End}	+		8.9	98.5
superscaffold00013 ^{End}	-	sgr02	0	22.4
superscaffold00036	-		22.5	33.1
gsc06 ^{End}	+		33.2	88.3
gsc15 ^{End End}	+	sgr03	0	76.5
superscaffold00091 ^{End}	+		0	1.7
superscaffold00065	-		1.8	6.5
gsc17	+	sgr04	6.6	22.5
superscaffold00088*	+		22.6	24.5
gsc17	+		24.6	36.3
superscaffold00095*	-		36.4	37.8
gsc17	+		37.9	41.2
superscaffold00081*	-		41.3	44.1
gsc17	+		44.2	62.1
superscaffold00085	-		62.2	64.4
superscaffold00051	-		64.5	71.1
superscaffold00087	-		71.2	73.4
superscaffold00094 ^{End}	+		73.5	75.0
gsc01	+		0	15.9
superscaffold00084*	-		16.0	18.3
gsc01 ^{End}	+		18.4	70.3
gsc05 ^{End}	+		0	48.4
superscaffold00059	-	sgr06	48.5	53.6
superscaffold00029	-		53.7	67.3
gsc11	+		0	44.9
superscaffold00060*	-	sgr07	45.0	50.1
gsc11	+		50.2	60.8
superscaffold00090	-		60.9	62.6
superscaffold00102 ^{End}	+		62.7	63.7

Sequence to assign		Target SGR		
Name	+/- ^a	Name	Start ^b	End ^b
gsc08 ^{End}	+	sgr08	0	57.7
gsc03 ^{End}	+	sgr09	0	57.4
gsc09	+	sgr10	0	5.0
superscaffold00048*	+		5.1	12.2
gsc09	+		12.3	57.2
gsc02 ^{End End}	+	sgr11	0	48.2
superscaffold00031	-	sgr12	0	13.0
gsc18	+		13.1	46.1
gsc13 ^{End}	+	sgr13	0	45.9
superscaffold00068 ^{End}	+	sgr14	0	4.5
gsc16	+		4.6	41.9
superscaffold00086	+		42.0	44.3
gsc12 ^{End End}	+	sgr15	0	41.9
gsc04	+	sgr16	0	9.6
superscaffold00089*	+		9.7	11.5
gsc04	+		11.6	38.4
superscaffold00099	-		38.5	39.6
gsc19	+	sgr17	0	35.0
superscaffold00077	+		35.1	38.3
gsc10 ^{End}	+	sgr18	0	37.2
gsc14 ^{End}	+	sgr19	0	10.6
superscaffold00105*	+		10.7	11.3
gsc14	+		11.4	17.9
superscaffold00097*	-		18.0	19.2
gsc14	+		19.3	24.2
superscaffold00098 ^{End}	-		24.3	25.5

^a Forward or reverse orientation; ^b The start and end coordinates are rounded to 0.1 Mb; ^{End} This GSC or superscaffold carries chromosome end annotation that was obtained by optical mapping; * This superscaffold was incorporated into an existing GSC.

3.1.2.6 Summary and Availability of the Genome Assembly

Comparing assembly A with E (Table 7) shows a reduction of the sequence number by a factor of 2.7 and an increase of the N50 length by a factor of 435. In each of the assembly steps, more bases were incorporated while the number of sequences decreased (Figure 5). The improved contiguity includes however an increase of the fraction of Ns from 9.9% to 31%.

Raw sequencing data (Table 2 and Table 3) as well as the final assembly E (Table 7) are deposited under the BioProject ID PRJEB5837 at the European Bioinformatics Institute. In parallel, a genome browser was set up and is maintained by the Genome Analysis Group at the FLI under <http://www.nothobranchius.info/NFINgb>. It is based on assembly E and includes alignments of selected sequence resources (e.g. BAC ends, genetic markers) as well as gene, repeat and variation annotations.

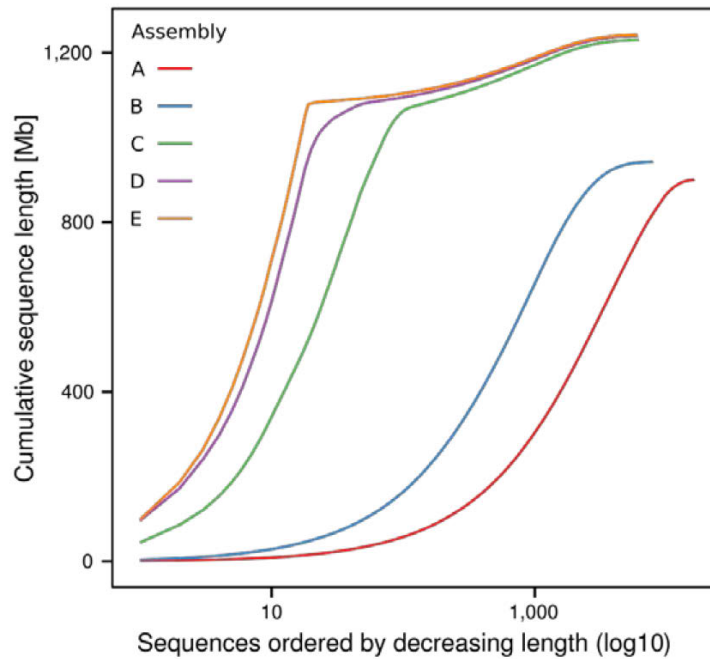


Figure 5: Cumulative Length of the Assemblies A-E.

The figure illustrates that the number of sequences decreases after each assembly step (right end of the curve) while less sequences are needed to reach a certain cumulative length (slope at the left side). Full name of the assemblies according to Table 7: A: ALLPATHS-LG (scaffolds), B: KILAPE + gap filling (scaffolds), C: Optical map integration, D: Genetic map integration, E: Comparative syntenic mapping. [modified from Figure S1C (Reichwald et al. 2015)]

3.2 Developing RepARK – A New Method for *de novo* Repeat Analysis

De novo repeat identification is an important step in genome analysis. There are only few tools adapted to high-throughput NGS data which are limited in applicability. RepARK was developed to provide a universal pipeline that creates species-specific repeat libraries based on unassembled NGS data.

3.2.1 The RepARK Pipeline

For the *de novo* repeat library creation, a set of NGS reads is analyzed for its k-mer content. Having the k-mers counted, their frequencies are transformed into a histogram that allows estimating which k-mers are derived from repeats and which belong to the unique fraction of the genome. Only the abundant k-mers are extracted and *de novo* assembled to form repeat consensi (Figure 6).

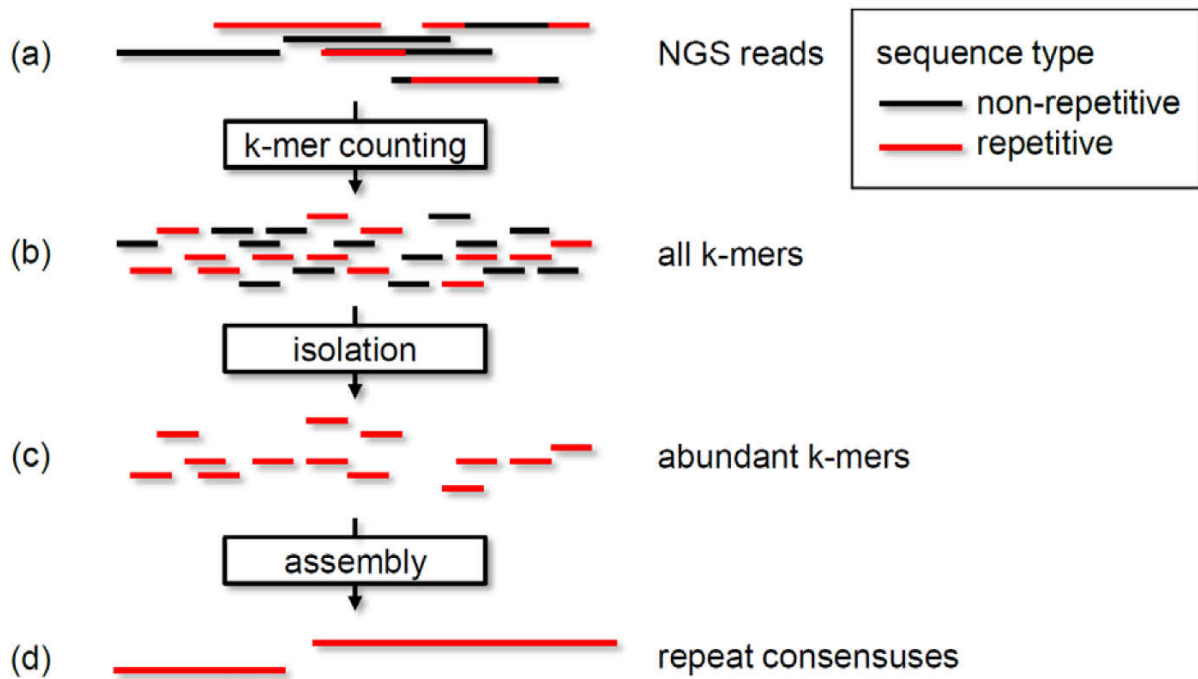


Figure 6: Outline of the RepARK Pipeline.

WGS NGS reads contain unique (black) and repetitive (red) fractions of the genome (a). K-mers of all reads (b) were identified and counted. Based on the count distribution the threshold of frequent k-mers is determined. Only k-mers that occur more often than this threshold are extracted (c). These abundant k-mers are subsequently assembled by a *de novo* genome assembly program (such as CLC or Velvet) to build repeat consensus sequences (d), which all together form the *de novo* repeat library. [from Figure 1 (Koch et al. 2014)]

3.2.2 Evaluation of RepARK Based on the *D. melanogaster* Genome

The evaluation of the performance of RepARK was carried out on the well-assembled and extensively studied genome of *D. melanogaster*. This genome was also chosen because of its high-quality repeat annotation that allows a comparison to the different repeat libraries built in this study.

3.2.2.1 *D. melanogaster* Repeat Library Construction

Two datasets of the *D. melanogaster* genome were obtained which contained 68 M simulated (“simulated”) or 83 M experimentally derived (“real”) reads. Of both datasets, 31-mers were extracted and counted. RepARK automatically determined the thresholds at which a 31-mer is considered repeat-derived to >60 and >84 for simulated and real data, respectively (Figure 7). The two assembly programs CLC and Velvet assembled the 2,675,416 simulated 31-mers and the 1,119,711 real 31-mers to four *de novo* RepARK repeat libraries. For comparison, I created additional repeat libraries using either RepeatScout and a Velvet *de novo* genome assembly or wgs-assembler with the respective simulated and real reads. Finally, *D. melanogaster* repeats from the database RepBase (“DmRepBase”) and a previously published library (“ReASLib”) were included in the following analyses (Table 11 and Table 12).

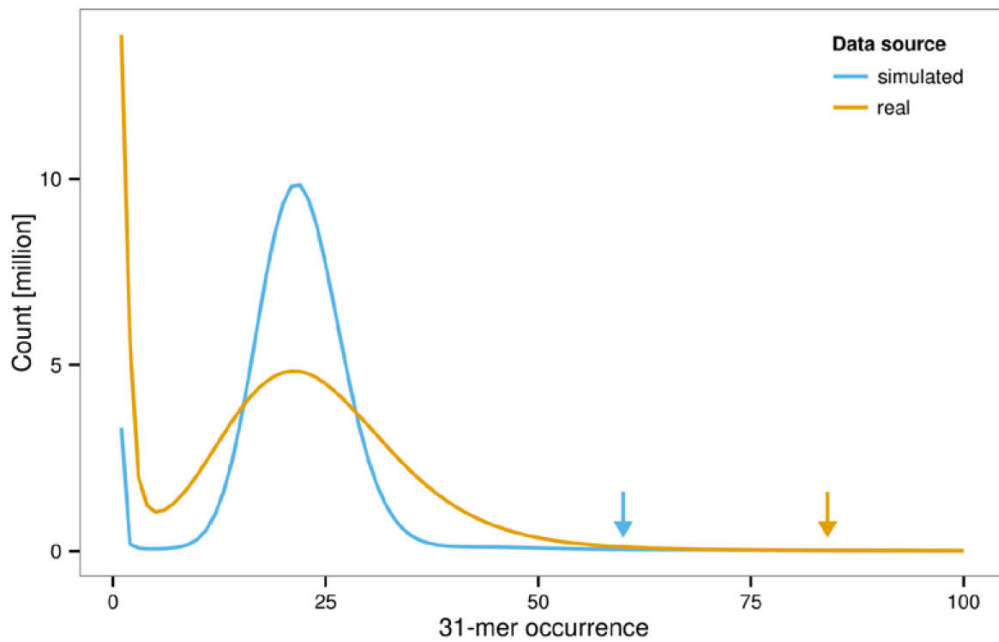


Figure 7: 31-Mer Coverage Histograms for Simulated and Real *D. melanogaster* Illumina Reads.

Arrows indicate the thresholds defining 31-mers as potentially repeat derived (simulated: >60, real: >84). [modified from Figure S1 (Koch et al. 2014)]

3.2.2.2 Basic Repeat Library Characteristics

When comparing the different repeat libraries, a general observation is the difference in number and length of repeat consensus sequences. This is primarily reflected by the N50 lengths of the repeat libraries generated by either RepARK, RepeatScout, or wgs-assembler which are one to two orders of magnitude (16-fold to 93-fold) smaller compared to either the RepBase or ReASLib repeat libraries. This indicates extensive fragmentation of the sequences within the repeat libraries. Additionally, the total length of libraries created by wgs-assembler and RepARK is much larger (2-fold to 7-fold) in respect to DmRepBase, which points to higher redundancy in these libraries. In terms of running time, the generation of RepARK libraries using either CLC (“RepARK CLC”) or Velvet (“RepARK Velvet”) was orders of magnitude (14-fold to 465-fold) faster than when using RepeatScout or wgs-assembler (Table 11 and Table 12).

Table 11: *D. melanogaster* Repeat Library Metrics from Simulated NGS Reads.

	RepeatScout	wgs-assembler	RepARK CLC	RepARK Velvet
Identification method	Velvet + RepeatScout	wgs-assembler surrogates	CLC	Velvet
Number of consensi	1,239	18,203	67,968	14,147
Total length [Mb]	0.174	4.3	4.3	1.9
Min/max length [bp]	51/2,565	66/6,446	30/6,945	57/6,943
N50 length [bp]	78	147	58	149
N90 length [bp]	64	116	36	59
Novel fraction [%]	36.7	29.3	33.1	31.1
Time to create [h]	8.75	284	0.61	0.61

[modified from Table 1 (Koch et al. 2014)]

Table 12: *D. melanogaster* Repeat Library Metrics from Real Data.

	DmRepBase	ReASLib	RepeatScout	wgs-assembler	RepARK CLC	RepARK Velvet
Source data	N/A	N/A	Sanger	Sanger	Illumina	Illumina
Identification method	Manual curation	Seed-based	Velvet + RepeatScout	wgs- assembler surrogates	CLC	Velvet
Number of consensi	249	391	414	14,296	19,677	4,284
Total length [Mb]	0.7	0.96	0.035	2.2	1.6	0.87
Min/max length [bp]	52/14,477	101/12,876	51/616	64/25,962	30/7,589	57/7,587
N50 length [bp]	5,402	4,757	83	158	87	290
N90 length [bp]	1,750	1,247	56	76	38	89
Novel fraction [%]	0.09	18.7	38.7	36.3	29.3	22.1
Time to create [h]	N/A	N/A	5.75	101	0.28	0.28

N/A: not applicable. [modified from Table 2 (Koch et al. 2014)]

3.2.2.3 Sensitivity

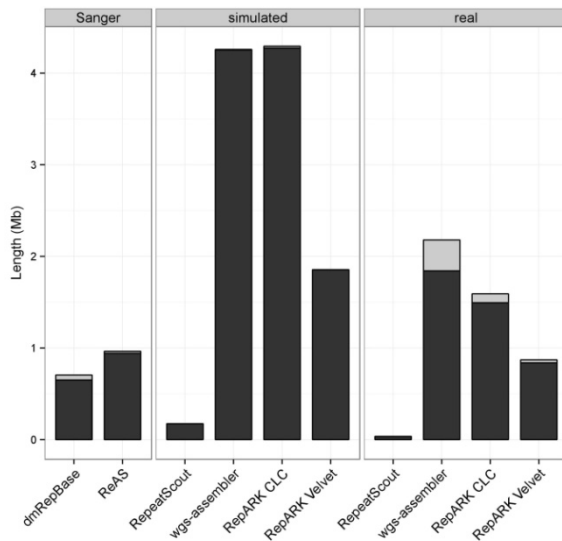
To check how much of the assembled libraries really represent repeats, each library was mapped onto the *D. melanogaster* genome using BLAT and filtered for a minimum identity of 80%. The major fraction of each repeat library (84-99%) fulfilled the requirement of mapping multiple times to the reference and was therefore considered “repetitive consensi” (Table 13 and Figure 8, black). The remaining sequences aligned only once or not at all (Figure 8, gray). This high proportion of repetitiveness is also observed when the identity threshold was set to 90% and 95% (Table 13). The library with the smallest fraction of repetitive sequences was built by wgs-assembler using real data which indicates the low sensitivity of this approach.

Table 13: BLAT Mappings of Repeat Libraries to the *D. melanogaster* Genome.

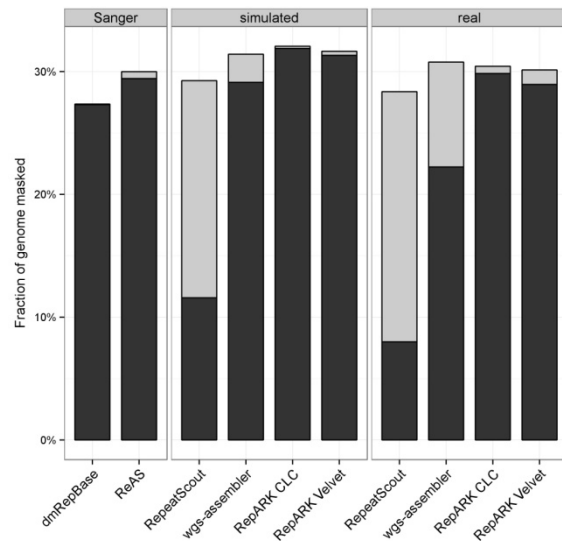
	Library	Number of hits at identity			Consensi with more than one hit at identity		
		>80%	>90%	>95%	>80%	>90%	>95%
Sanger	DmRepBase	20,627	17,241	12,302	213 (86%) ^a	212 (85%)	209 (84%)
	ReASLib	38,734	34,539	23,639	342 (87%)	342 (87%)	341 (87%)
Simulated reads	RepeatScout	50,609	47,897	29,857	1,236 (100%)	1,234 (100%)	1,207 (97%)
	wgs-assembler	1.5 M	1.4 M	1.0 M	18,165 (100%)	18,164 (100%)	18,133 (100%)
	RepARK CLC	3.4 M	3.3 M	2.9 M	67,419 (99%)	67,418 (99%)	67,407 (99%)
	RepARK Velvet	0.7 M	0.7 M	0.5 M	14,141 (100%)	14,141 (100%)	14,141 (100%)
Real reads	RepeatScout	19,147	18,330	11,447	411 (99%)	409 (99%)	391 (94%)
	wgs-assembler	1.1 M	1.1 M	0.9 M	10,627 (74%)	10,625 (74%)	10,571 (74%)
	RepARK CLC	0.9 M	0.8 M	0.7 M	18,093 (92%)	18,088 (92%)	18,047 (92%)
	RepARK Velvet	0.2 M	0.2 M	0.2 M	4,050 (95%)	4,048 (94%)	4,035 (94%)

^a Percentages refer to the total number of consensi within a library. [modified from Table S2 (Koch et al. 2014)]

The libraries from RepeatScout were almost entirely composed of repetitive sequences (Figure 8), but their total length was significantly smaller compared to the remaining libraries (simulated: 174 kb, real: 35 kb, see Table 11 and Table 12). To investigate the reason for this, the repeat content of the underlying Velvet genome assemblies was analyzed with RepeatMasker and the DmRepBase library. Of the Velvet assemblies of either simulated or real data, 6.5% and 4.7% was found by RepeatMasker, respectively, which is at most one quarter of the DmRepBase-repeat content in the *D. melanogaster* reference assembly (27%, Figure 9).

**Figure 8: Total Length of the *D. melanogaster* Repeat Libraries.**

Repetitive bases mapping multiple times to the reference genome and are depicted in black whereas non-repetitive bases are gray. [modified from Figure 2 (Koch et al. 2014)]

**Figure 9: Repetitive Genomic Fraction of the *D. melanogaster* Reference Genome.**

Repeats that are masked with the respective library are depicted in black whereas additional repeats from RepBase are gray. [modified from Figure 3 (Koch et al. 2014)]

The fraction of the reference genome considered repetitive by the several repeat detection approaches was determined by RepeatMasker using the corresponding library. More reference sequence was identified as repetitive when either the RepARK libraries or ReASLib was used compared to when using RepBase. Of state-of-the-art methods, wgs-assembler-based repeat libraries provided comparable results only for simulated reads while the two RepeatScout derived libraries masked only a small fraction of the reference (Figure 9, black). This result was to be expected based on the small library size. To see which repeat fraction was missing, another repeat masking round of the already masked reference was done with the “gold standard” DmRepBase. Only a small fraction of the genome sequence masked using DmRepBase was not identified as repetitive when using RepARK libraries (0.18-1.18%) and ReASLib (0.56%) (Figure 9, gray). In contrast, wgs-assembler (2.3-8.5%) and RepeatScout (17-20%) derived libraries left much more unmasked.

Next I analyzed whether the known and well established repeats from RepBase are represented in the *de novo* libraries. This representation is called “completeness” and is indicated by the fraction of an element’s bases that are also found in *de novo* libraries. In Figure 10, 212 TEs of DmRepBase were grouped according to their major classes into nonLTR (41), LTR (138) and DNA (33). The remaining 37 elements from DmRepBase were not considered as they are marked either “Unknown”, “Simple”, “Low Complexity”, “ARTEFACT” or “RNA”. In general, LTR and nonLTR retrotransposons showed a higher median completeness than DNA transposons. However, RepARK libraries consistently showed as good or superior completeness compared to the other libraries investigated. Details of RepARK library completeness including all 249 DmRepBase repeats are shown in Figure 11.

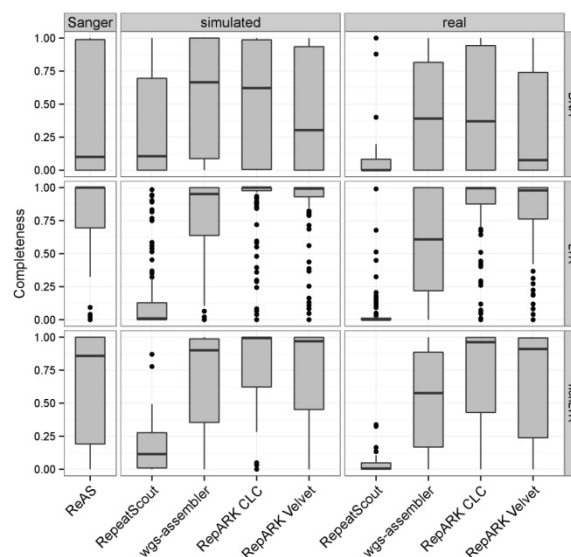


Figure 10: Completeness of 212 *D. melanogaster* RepBase Repeats Separated by their Main TE Class.

Box: first and third quartiles; horizontal line: median; whiskers: most extreme value within 1.5-fold of inter-quartile range; dots: outliers. [from Figure 4 (Koch et al. 2014)]

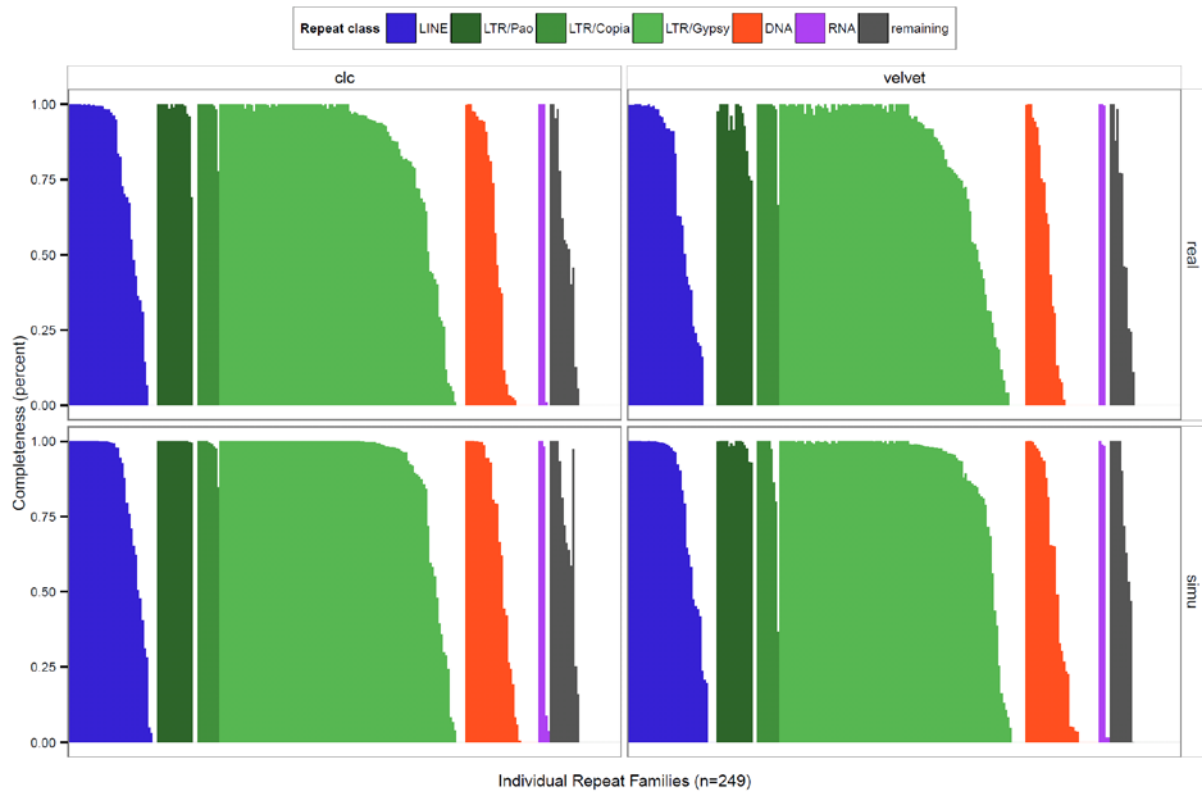


Figure 11: Completeness of the 249 *D. melanogaster* RepBase Repeats in the Four RepARK Libraries.

The bars represent individual DmRepBase consensi and are ordered by their mean that was calculated over all four analyzed libraries. Empty regions at the right side of each class represent consensi with zero completeness.

3.2.2.4 Identification of Putative Novel Repeats

Previously the genomic fraction of RepBase identified repeats that were not found by the *de novo* libraries was determined (Figure 9, gray). However, these libraries might also contain repeat consensi that are not yet represented in RepBase and therefore might be novel. Therefore, each library was repeat-masked with DmRepBase leaving only those consensi unmasked that are novel (ranging in the RepARK libraries from 22.1-33.1%, see Table 11 and Table 12). These consensi were then mapped to the *D. melanogaster* reference genome. By this approach, consensi with different mapping patterns were found. One fraction mapped with high identity proximal to one another on the same chromosome (Supplemental Figure 3) and/or to the corresponding heterochromatin entry (Supplemental Figure 4). These patterns reflect the characteristics of SDs. To assess these findings, the affected regions were compared to a list of known SDs of *D. melanogaster*. The largest fraction of the SDs could be identified by the RepARK libraries compared to the other *de novo* repeat libraries studied (Figure 12), with the exception of wgs-assembler surrogates using simulated data.

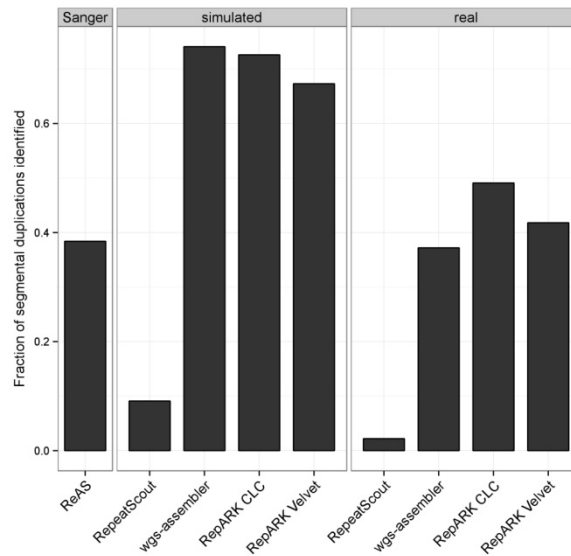


Figure 12: Genomic Fraction of Known SDs and their Representation in the Repeat Libraries.
[from Figure 5 (Koch et al. 2014)]

3.2.2.5 Classification of *D. melanogaster* Repeats

To analyze the composition of the libraries in respect to the repeat classes, TEclass was first applied to all consensi ≥ 50 bp. In each library, TEclass successfully classified more than 90%. Only the top-level classifications into class I and class II TEs were evaluated. A greater proportion of RepARK consensi were classified as DNA transposons (class II) and a fewer proportion as retrotransposons (class I), compared to ReASLib or DmRepBase (Supplemental Table 3). The opposite was observed when using these classified consensi for masking the *D. melanogaster* reference genome sequence; there, more DNA transposons and less retrotransposons were identified with the RepARK libraries than with the DmRepBase annotation (Figure 13A). This bias could be due to the extensive fragmentation of the RepARK libraries to which the TEclass algorithm may not be adapted. Thus, in a second classification attempt TEclass was restricted to consensi >100 bp which considerably reduced the bias toward DNA transposons in the repeat annotation of the *D. melanogaster* genome using RepARK libraries (Figure 13B).

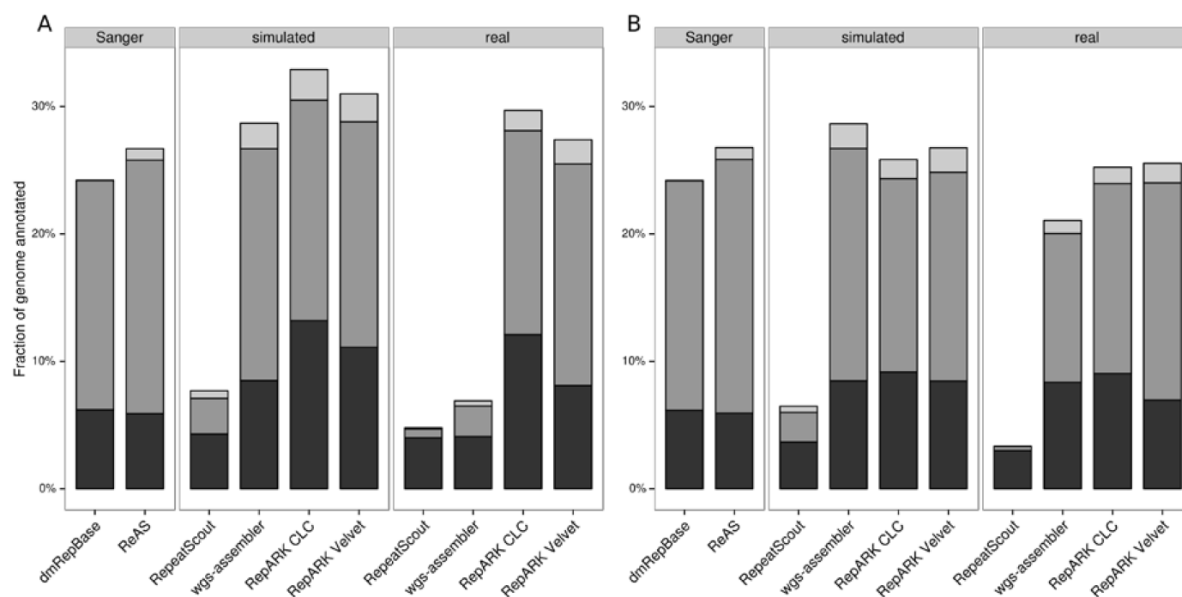


Figure 13: Genomic Fraction of TEclass-Classified Repeat Consensi.

The repeat consensus classification was either DNA transposon (black), retrotransposon (dark gray) or unclear (light gray). The repeat libraries either contain consensi (A) ≥ 50 bp or (B) > 100 bp. [modified from Table S5 and Figure 6 (Koch et al. 2014)]

3.2.3 RepARK on the Complex Human Genome

3.2.3.1 Human Repeat Library Construction

Furthermore, RepARK was applied to the much larger and more complex human genome. For this, 1,307 M Illumina reads with 40-fold genome coverage were downloaded from the SRA. RepARK determined a repetitive 31-mer threshold of 76 and built a repeat library with 62,425 consensi from 28,481,680 31-mers using Velvet (Table 14). Comparing this library to the gold-standard human RepBase library (“HsRepBase”), the RepARK library was substantially longer (7.9 Mb vs. 1.6 Mb for HsRepBase). Of the human RepARK library, 93% was found to be repetitive which is consistent with the findings in *D. melanogaster* (Figure 9).

Table 14: Human Repeat Library Metrics and Mapping Results Against the Human Reference Sequence.

	HsRepBase	RepARK Velvet
Number of consensi	1,439	62,425
Total length [kb]	1,566	7,882
Min/max length [bp]	63/9,044	57/42,518
N50 length [bp]	2,822	143
N90 length [bp]	471	57
Time to create [h]	N/A	22
Number of consensi with multiple hits	1,167 (81% ^a)	57,239 (92% ^a)
Total length of consensi with multiple hits [kb]	1,471 (94% ^b)	7,318 (93% ^b)

^aRatio to the total number of consensi of the library. ^bRatio to the total length of the library, N/A: not applicable. [from Table 3 (Koch et al. 2014)]

3.2.3.2 Detection of the Epstein-Barr Virus Genome in Human Data

Analyzing the human dataset revealed an unexpected additional feature of the RepARK approach. Velvet assembled a number of very long consensi from these data, with the longest being 42,518 bp long which is almost twice as long as the longest known LTR retrotransposon *ogre* with 25 kb (Macas and Neumann 2007). Comparing this consensus with the “Nucleotide collection (nt/nr)” from NCBI¹⁷, a highly significant match to the Epstein-Barr virus (EBV alias Human herpes virus 4) was identified. This virus had been used to establish the human cell line Coriell (Coriell Institute, GM12878) which had then been sequenced at the Broad Institute (BioProject ID PRJNA52009). Consequently, the entire human RepARK library was aligned to the EBV reference genome. In total, 23 repeat consensi (135 kb) with >90% coverage and a p-value < 10⁻⁶⁰ were identified. Together, these consensi have 99.58% identity to the virus genome. The majority (90.5%) of the 171 kb virus genome is covered by at least one consensus (Figure 14).

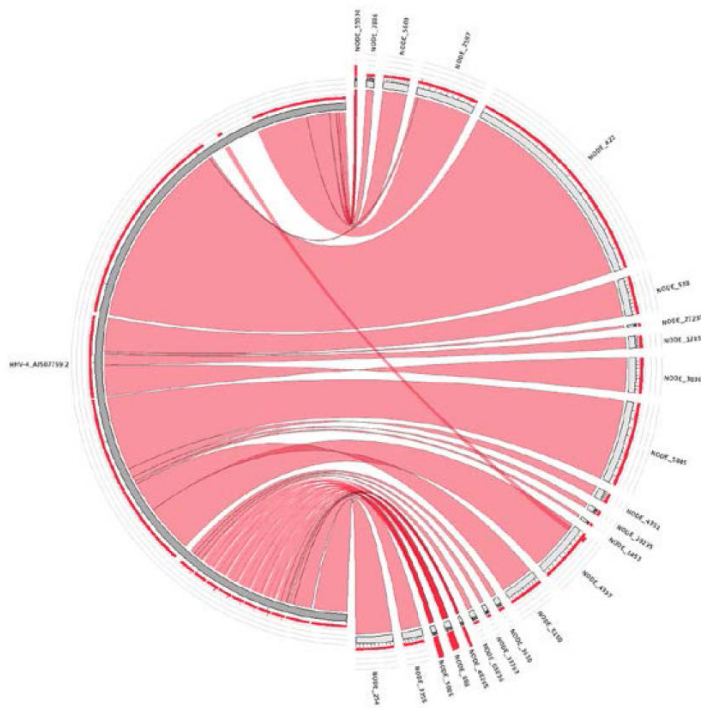


Figure 14: High Confidence Alignments of Human RepARK Consensi to the EBV Genome.

On the left side is the EBV genome and on the right side RepARK consensi. Each red ribbon represents a consensus alignment with >90% mapping and $p < 10^{-60}$, encompassing 90.5% of the EBV genome. Lower confidence consensi align to the remaining 9.5% with more relaxed criteria. Three consensi map multiple times to the virus genome sequence (NODE_48265, NODE_888, NODE_5085; dark red). The graphic was created with Circoletto (<http://bat.ina.certh.gr/tools/circoletto/>). [from Figure 7 (Koch et al. 2014)]

Furthermore, it was possible to estimate the copy number of EBV genomes within the analyzed dataset. For this, the set of 31-mers used for the repeat library assembly was mapped with bowtie (Langmead et al. 2009) onto the previously identified 23 consensi. In total, 141,486 31-mers mapped to the consensi showing a mean k-mer count of 690 (referring to the peak in Figure 15) while 93.4% have a frequency above 200. Assuming no bias in sequencing and calculating the ration of mean frequency and coverage, the EBV genome may occur in 17 copies per haploid genome in the analyzed 40-fold dataset.

¹⁷ <http://blast.ncbi.nlm.nih.gov>

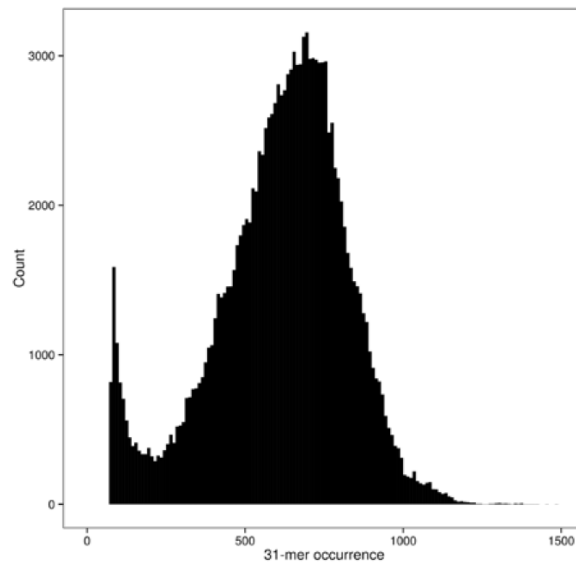


Figure 15: Distribution of the Occurrence of 31-mers that Map to the EBV Genome.
Occurrence refers to the number of a particular 31-mer in the entire human dataset.

3.3 Comprehensive *N. furzeri* Repeat Analysis and Genome Annotation

This section describes the generation of an as complete as possible library of the *N. furzeri* genomic repeats. This library is a prerequisite for many downstream-analyses of the assembled genome sequence, such as gene prediction or detection and analysis of sequence variations. Moreover, knowing the repetitive fraction of the *N. furzeri* genome is necessary for understanding its biology and evolution as well as for future efforts aimed at improving the reference sequence quality in terms of completeness and contiguity. I used multiple approaches to benefit from their advantages and to compensate for their drawbacks. I applied established methods to both Sanger read data and the genome assembly as well as the newly developed RepARK (Koch et al. 2014) program to NGS reads. I then used the combined repeat library to identify and annotate the final *N. furzeri* genome assembly E (Table 7, in the following referred to as “genome assembly”).

3.3.1 Basic Repeat Analysis of the Genome Assembly

For subsequent comparisons, a first insight into the repeat composition of the genome assembly was obtained by applying RepeatMasker with the built-in library of common vertebrate repeats (version 20150807, “-species vertebrates”). This analysis identified 12.96% of the 1.24 Gb genome assembly as repetitive.

3.3.2 Building a Library of *N. furzeri* Repeat Consensi

Three *N. furzeri* repeat libraries were created by using different strategies: First, RepeatModeler was applied to the genome assembly, resulting in 1,064 (0.7 Mb) repeat consensus sequences (“RModLib”). Second, prior to this thesis, a library was build from a sample of 120 Mb WGS Sanger sequences (Table 2) using RepeatMasker with RepBase followed by RepeatScout (4,859

sequences; 5.3 Mb; “SangerLib”) (Koch 2010). Of this library, only unidentified sequences (labeled as ‘Unknown’; 3,386, 1.0 Mb) were included in the following analyses. RepeatScout identified most of these “unknown” consensi and they were therefore considered as novel *N. furzeri*-specific repeats while those found by RepeatMasker and RepBase are already known in other species. The third repeat library was created based on 47.5 Gb WGS Illumina sequences using the RepARK pipeline (“RepARKLib”). More than 1.5 billion different 31-mers were counted and a cut-off threshold of 66 was automatically determined. The 22.4 M 31-mers occurring at least 67 times were assembled to 248,033 repeat consensi (17.9 Mb; Table 15, top).

From each of the three libraries, consensi shorter than 80 bp were removed and SangerLib as well as RepARKLib were subjected to redundancy reduction based on their similarity (Table 15, bottom). The remaining consensi built a combined set of 33,028 sequences and were subjected to another round of redundancy removal. This resulted in the redundancy-free CombinedLib which comprises 24,954 (5.6 Mb) consensus sequences.

Table 15: Repeat Libraries of *N. furzeri*

Library name	RModLib	SangerLib	RepARKLib	CombinedLib
Underlying data	Genome assembly	Sanger sequences	Illumina reads	Combination
Method	RepeatModeler	TRF, RM, RS	RepARK	CD-HIT-EST
Initial situation				
Total length [Mb]	0.751	5.3	17.9	-
Min / max length [bp]	35 / 6,342	51 / 32,616	50 / 5,111	-
N50 length [bp]	1,112	4,606	58	-
N90 length [bp]	354	428	53	-
Repeat consensi	1,064	4,859	248,033	-
Classified	467	1,473	0	-
Not classified	597	3,386	248,033	-
After discarding consensi shorter 80 bp and redundancy reduction				
Total length [Mb]	0.748	0.855	5.120	5.613
Min / max length [bp]	81 / 6,342	80 / 13,972	80 / 5,111	80 / 13,972
N50 [bp]	1,126	526	184	300
N90 [bp]	365	157	88	98
Repeat consensi	991	2,397	29,640	24,954
Classified	458	0 ^a	0	458
Not classified	533	2,397	29,640	24,496
After classification of the CombinedLib				
Classified	-	-	-	7,425
Not classified	-	-	-	17,529

^aThe 1,473 classified consensi from the input SangerLib were excluded from further steps (see text); RM: RepeatMasker; RS: RepeatScout.

3.3.3 Classifying Consensi of the Combined Repeat Library

Most sequences of the CombinedLib were not classified (24,496/98%; N50 265 bp; total length 5.17 Mb/92%; Table 15) meaning that the class or family they belong to was unknown. I attempted their classification applying a multi-step approach (Figure 16). After the first iteration of Censor using RepBase reference repeats, 4,509 of the unknown consensi were classified and in a second turn in which the FishRepLib (see chapter 2.3.4) was passed to Censor, another 2,228 sequences were classified. Of these, eleven repeat consensi were classified as tRNA-related and 173 as MITEs. Another 38 consensi were identified as TRs due to their high similarity to the 77-bp and 348-bp *N. furzeri*-specific TRs. In the last steps, including manual inspection, the identification of two Helitrons and six AT-rich elements finalized the classification of the CombinedLib which contained 7,425 classified (29.8%; N50 373 bp; total length 1.89 Mb/33.7%) and 17,529 unclassified (70.2%; 271 bp; 3.72 Mb/66.3%) repeat consensi (Table 15).

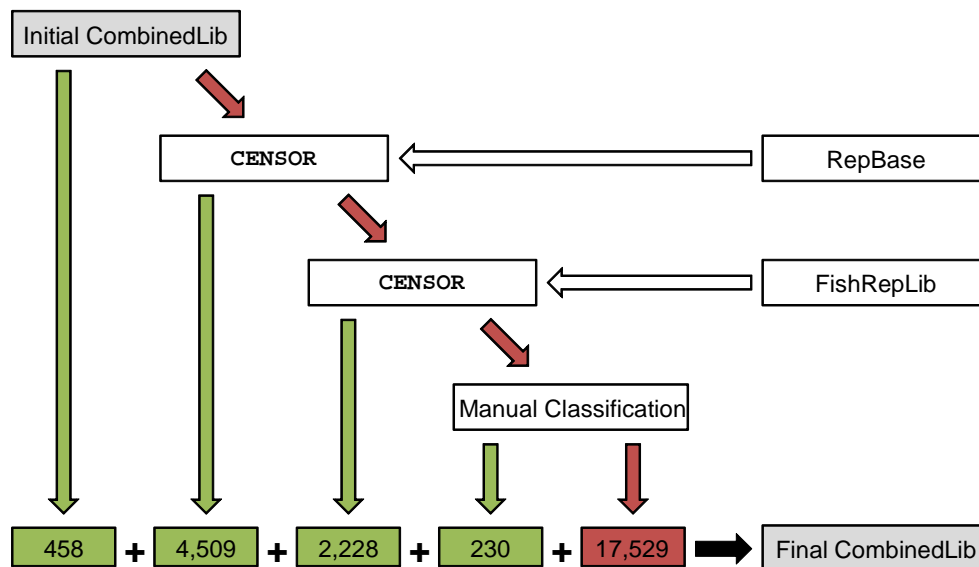


Figure 16: Repeat Classification of the CombinedLib.

The initial CombinedLib contains 24,954 of which 458 are already classified. Three steps of classification were applied while after each successfully classified consensi were excluded from the input of the subsequent step (green). A total of 7,425 consensi were classified whereas 17,529 remain unknown (red).

3.3.4 Repeat Annotation of the Genome Assembly

3.3.4.1 Performance of Single Repeat Libraries

To compare the single repeat libraries, i.e. libraries created by applying one approach only, and the CombinedLib, they were separately applied to determine the repetitive fraction of the genome assembly (Table 16). Of the individual libraries, the RModLib annotated with 33.1% the highest fraction of the genome assembly while the SangerLib and RepARKLib annotate with 22.9% and 26.0% a smaller fraction. Using the CombinedLib, 35.1% of the genome assembly was masked as repetitive.

Table 16: Repeat Fraction of the Genome Assembly Annotated by Different Repeat Libraries of *N. furzeri*

Genome assembly	RModLib	SangerLib	RepARKLib	CombinedLib
With Ns (1.24 Gb)	33.14%	22.94%	26.01%	35.10%
Without Ns (0.89 Gb)	48.06%	33.26%	37.71%	50.90%

3.3.4.2 Repeat Composition of the Genome Assembly

A detailed repeat analysis was conducted to discover the different classes of repeats in the genome assembly. For this, TRF was first applied and detected 2.07% as TRs and masked these with Ns. Based on this, RepeatMasker using the CombinedLib determined an additional 33.41% of the genome as repetitive which summed up the total repeat content to 35.48% (Table 17). Among TRs, microsatellites occurred twice as often as minisatellites (156,901 vs. 86,445) while satellites were much less represented (8,833). Although nearly half of the DRs were unclassified (“Unknown”), the predominant class was LINE elements (8.4%), followed by DNA (5.8%), LTR (1.9%) and SINE elements (1%).

Table 17: Repeat Composition of the *N. furzeri* Genome Assembly

Repeat type	Class / type	Count	Masked bases	Assembly fraction
Tandem repeats^a	microsatellites	156,901	11,682,565	0.94%
	minisatellites	86,445	12,424,246	0.99%
	satellites	8,833	5,118,661	0.41%
	Total Tandem repeats	252,179	25,736,111	2.07%
Dispersed repeats	DNA	240,198	71,864,559	5.78%
	LTR	57,503	23,036,236	1.86%
	SINE	69,713	12,153,172	0.98%
	LINE	270,688	104,225,372	8.39%
	Simple	14,048	5,108,876	0.41%
	Unknown	828,055	191,000,684	15.37%
	Total Dispersed	1,479,862	407,343,887	32.79%
Others	AT_rich	63	3,312	0.00%
	Vector	25	2,409	0.00%
	Low complexity	142,110	7,739,848	0.62%
	Total Others	142,198	7,745,569	0.62%
Total		1,622,403	440,870,579	35.48%^b

^a all values of the TR section refer to merged intervals of overlapping TRs (italic) except the total number of bases and the total genome assembly fraction which are based on actual counts of Ns within the genome assembly. ^b used for genome assembly completeness calculation in chapter 3.3.6 as the upper bound repeat value.

3.3.4.3 Evolutionary History of Transposable Elements

Analyzing the difference between the repeat copies in the *N. furzeri* genome allows an estimation of the evolutionary history of repetitive elements. The repeats found in the genome assembly were analyzed for their Kimura distances which were then cumulatively plotted as histogram allowing the identification of transpositional bursts (Figure 17). In *N. furzeri*, two strong bursts of transposition are observed when combining all four major repeat classes: (i) a recent burst with a peak at Kimura distance 4 and (ii) an older at a distance of 40.

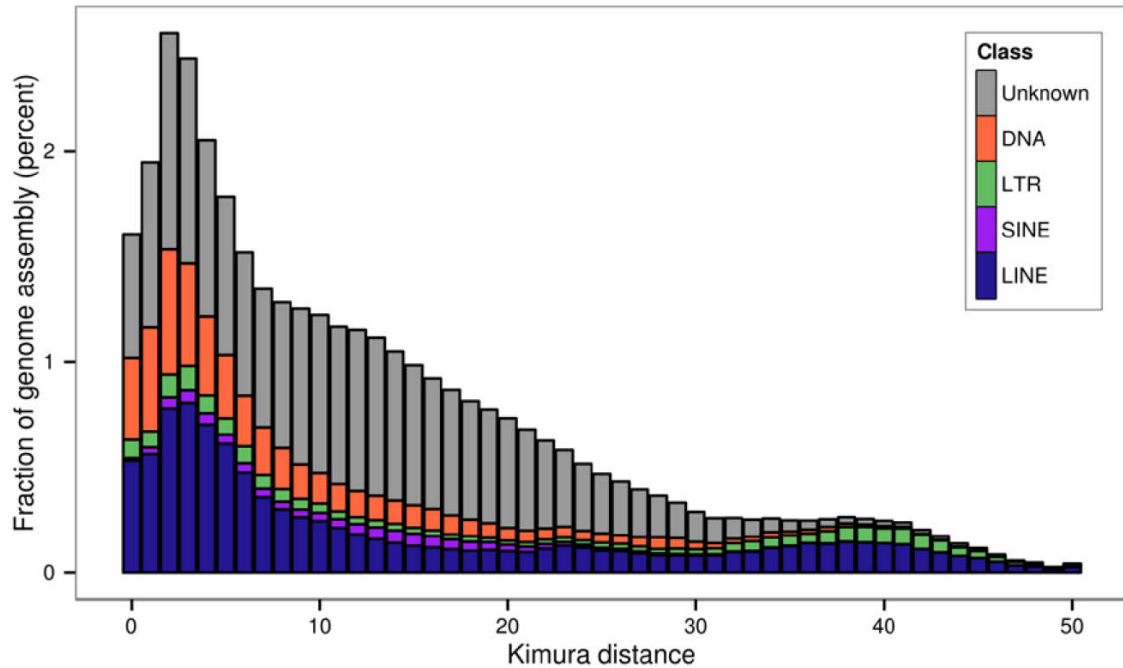


Figure 17: Evolutionary History of the Major TE Classes in the *N. furzeri* Genome Assembly.

Only superfamilies that cover at least 0.2% of the genome assembly are added to their respective classes. Column “0” contains all repeats with a Kimura distance of <1 while category “50” contains repeats with a distance of ≥ 50 . The group of repeats that could not be classified are referred to as “Unknown”.

When the repeat classes were plotted individually, a much more complex picture of the evolutionary history of TEs emerged (Figure 18). DNA elements represented (after LINES) the second largest class of TEs present in the *N. furzeri* genome assembly (5.8%) and clearly showed a relatively young peak (Kimura distance 0 to 3). This peak was mainly composed of hAT and TcMar elements while hAT elements are the most abundant DNA transposons. A minor peak of TcMar elements was also observed at a Kimura distance of 28. LTRs were predominantly represented by Gypsy elements that were nearly exclusively part of a relatively old peak at Kimura distance 39. They also contributed to a younger and more diverse peak which additionally contained Ngaro, ERV and DIRS elements. Although BEL-Pao is one of the genome-wide poorly-represented LTR superfamilies, it shows a quite unique profile with three peaks of at Kimura distances 43 (extremely old), 13 (intermediate) or 0 (very young). SINEs were only represented by 1% in the genome assembly and together formed two peaks of which one is relatively young (Kimura distance 1 to 4) while the second is of an intermediate age (Kimura distance 14). MIR elements are the most frequent representatives among SINEs (0.6%). For LINES, a strong and very young peak was mainly composed of RTE and L2 elements (Kimura distance 0 to 3) while a smaller and older peak largely contained L2 or Rex-Babar elements (Kimura distance 38). The L2 superfamily was also the most frequent of all superfamilies and occupied in total 4.8% of the genome assembly. Taken together, LINE, SINE and LTR elements often showed two clear peaks in their profiles while DNA elements have been active only recently. The profile of the “unknown” elements showed similarities to the profiles of all four TE classes but with a clear absence of the very old burst which was observed for LINE and LTR elements.

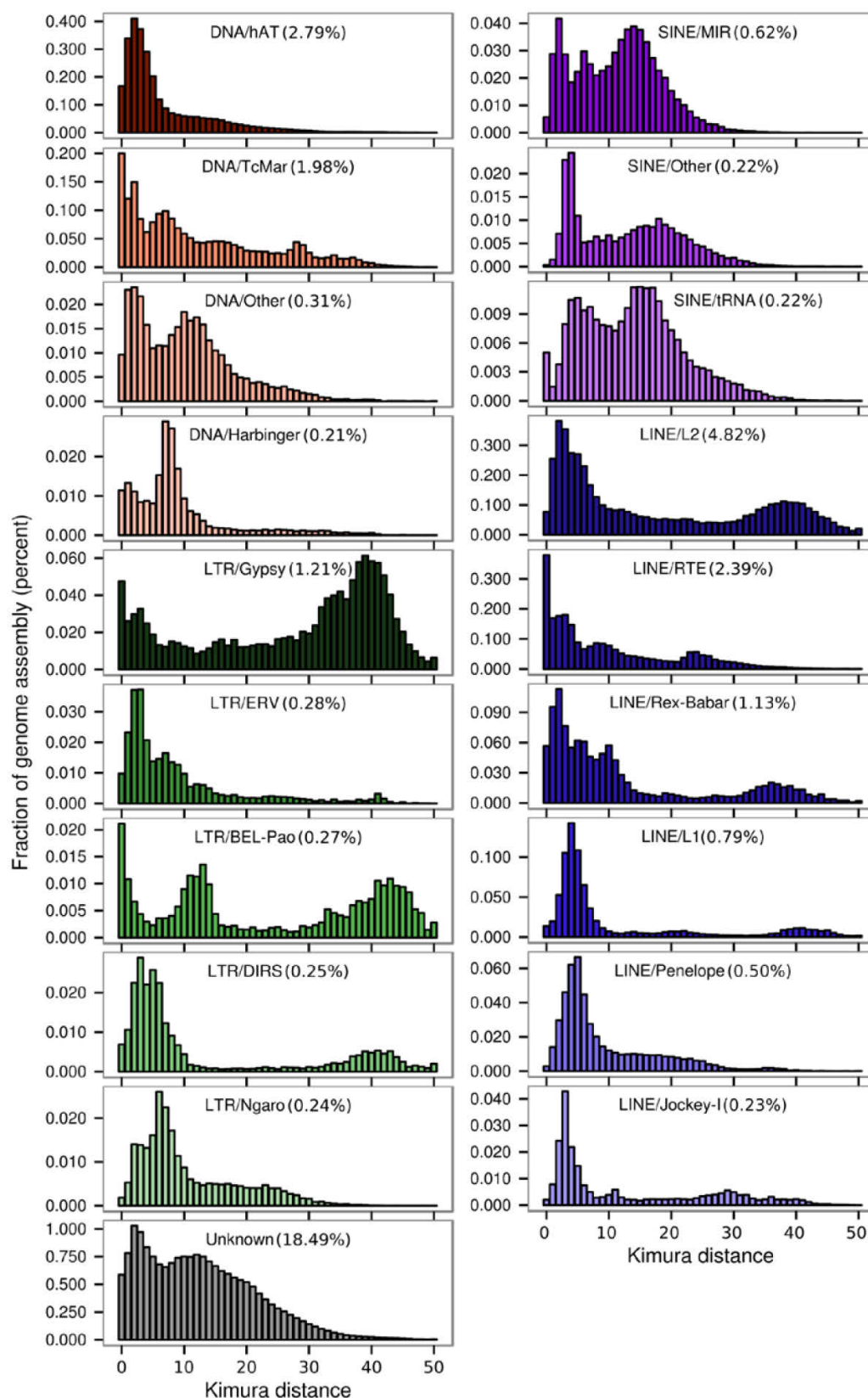


Figure 18: Evolutionary History of TE superfamilies in the *N. furzeri* Genome Assembly.

All superfamilies that cover at least 0.2% of the genome assembly are displayed. The genomic fraction is given after each superfamily name. Column “0” contains all repeats with a Kimura distance of <1 while category “50” contains repeats with a distance of ≥ 50 . The group of repeats that could not be classified are referred to as “Unknown”. Note, that the y-axes are in different scale but always display the percentage of genome assembly.

3.3.5 Determination of the Repetitive Fraction of the *N. furzeri* Genome

As shown above, the genome assembly contained about 35.5% repeats. However, this is an underestimate of the repeat content of the *N. furzeri* genome (Reichwald et al. 2009) and most likely due to collapsing of repetitive structures during its assembly. Therefore, I analyzed the repeat content of unassembled Sanger, 454, Illumina and PacBio sequencing reads. Upper and lower bounds of estimated repeat contents were calculated with two different strategies: (i) TRF was used to detect and mask TRs followed by a two-pass DR detection step with RepeatMasker and RepBase followed by *de novo* DR detection with RepeatScout. (ii) TRs were identified and masked by RepeatMasker and a repeat library comprised of the two most abundant TRs (a G+C-rich 77 bp minisatellite and a G+C-poor 348 bp satellite (Reichwald et al. 2009)) followed by a one-pass DR detection using RepeatMasker and the CombinedLib. Both strategies were applied to the reference sequence, the Sanger WGS reads and the PacBio ROIs; here, respective sequence lengths were sufficient for *de novo* repeat detection by TRF and RepeatScout. To account for the short length of the Illumina and 454 reads, only the second strategy, which does not involve a *de novo* detection step, was applied and both datasets were sub sampled to fractions of 0.1%. Comparing all four datasets, the lower bound repeat estimate was at 55.6% (454 data analyzed with the second approach) while the upper bound estimate was at 70.2% (Sanger data analyzed with the second approach) (Figure 19). These results support previous repeat estimates and show that the genome assembly indeed lacks a substantial fraction of repeats.

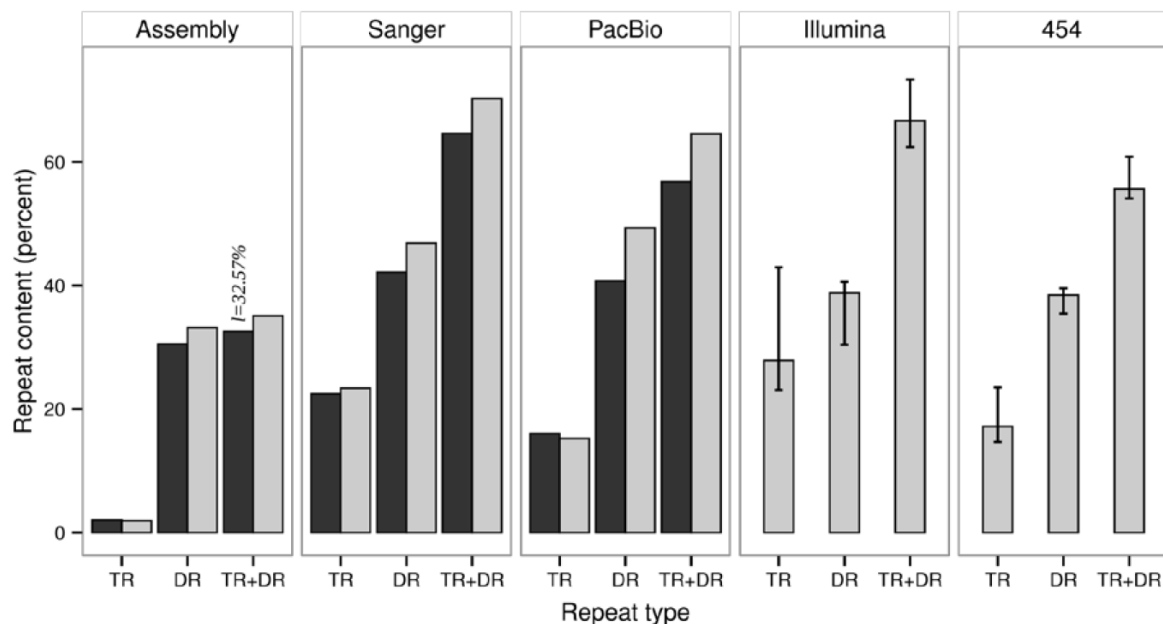


Figure 19: Repeat Content in Assembled and Non-Assembled *N. furzeri* Data.

The reference sequence (Assembly) as well as Sanger WGS reads and PacBio ROIs were analyzed by two approaches (described in the main text) for obtaining lower-bound (black) and upper-bound (gray) estimates of the repeat content. For Illumina and 454 reads, 0.1% fractions of each dataset or sequence run were randomly sampled and analyzed. Error bars show the minimum and maximum of all samples. Note that Illumina and 454 reads could be analyzed only using the upper-bound approach due to their short length. Tandem repeat (TR) and dispersed repeat content (DR) is shown separately as well as the sum of both (TR+DR). The value which is marked with “ $l = 32.57\%$ ” is used for genome assembly completeness calculation in chapter 3.3.6 as the lower bound repeat content value. [modified from Figure S2B (Reichwald et al. 2015)]

3.3.6 Repeat-Based Estimation of the Completeness of the *N. furzeri* Genome Assembly

The current genome assembly has two shortcomings: (i) with a total length of 1.24 Gb it is considerably shorter than its experimentally estimated size of 1.5 Gb (Reichwald et al. 2009), and (ii) the 1.24 Gb assembly contains 31% of Ns (introduced in the scaffolding process). Therefore, it is of prime interest to find out to which extent non-repetitive and repetitive sequences are present or missing in the genome assembly as well as to identify means by which the assembly could be improved further.

To estimate the non-repetitive fraction of the genome assembly (i.e. assembly completeness), the repeat content of sequence gaps with respect to their lengths was evaluated. Accordingly, gap-spanning sequences were obtained by running PBJelly with PacBio WGS sequences and by aligning assembled PacBio BAC-derived contigs to the reference sequence. PBJelly filled 19,065 gaps of ≥ 25 nt (16.2 Mb) and the BAC contig alignments filled another 254 gap sequences (509 kb). In those, TR and DR content was determined at 83.3% in total and a higher fraction of repeats was observed for longer gaps (Table 18). A logarithmic function ($y = 15.66 \ln(x) + 60.10$) was fitted to the average repeat fraction within different gap lengths and allowed an estimation of a repeat content of $>92.6\%$ in gaps longer than 72.9 kb.

When all gaps of the genome assembly are categorized into length bins like done for the filled gaps, the repetitive and unique bases can be estimated from that approximation as well (Table 19). These fractions were calculated using the approach described in chapter 2.3.6 where $G = 1,500,000,000$ denotes the *N. furzeri* genome size (Reichwald et al. 2009), $A = 1,242,498,532$ and $N_A = 385,709,457$ are given in Table 7, $X_g = 94.51\%$ is calculated in Table 18, $R_g = 351,784,276$ and $U_g = 33,925,181$ are given in Table 19. Based on a lower ($l = 32.57\%$, given in Figure 19) and an upper ($u = 35.48\%$, given in Table 17) repeat content estimate, I calculated that between 90.39% and 89.64% of the unique fraction of the *N. furzeri* genome is contained within the genome assembly. Accordingly, only 40.48% to 42.55% of the entire repetitive fraction of the *N. furzeri* genome is represented in the current genome assembly.

Table 18: Filled Gaps Using PacBio Sequences.

Gap class	Size range [bp]	Gap number	Overall size [bp]	Repetitive [bp]	Repeat ratio	Approximation ^a
1	1-99	5,106	191,472	108,233	56.53%	60.10%
2	100-299	3,169	605,292	452,756	74.80%	70.96%
3	300-899	5,375	3,069,881	2,413,379	78.61%	77.31%
4	900-2699	4,350	6,343,982	5,250,580	82.76%	81.81%
5	2,700-8,099	1,200	5,236,145	4,613,288	88.10%	85.31%
6	8,100-24,299	116	1,181,998	1,009,068	85.37%	88.16%
7	24,300-72,899	3	87,136	76,734	88.06%	90.58%
8	72,900-218,699					92.67%
9	≥218,700					94.51%

^ay=15.66ln(x)+60.10. [modified from Table S1P (Reichwald et al. 2015)]

Table 19: Gaps of the Genome Assembly.

Gap class	Size range [bp]	Gap number	Overall size [bp]	Approximation	
				Repetitive [bp]	Unique [bp]
1	1-99	28,115	594,616	357,388	237,228
2	100-299	9,817	1,826,104	1,295,777	530,327
3	300-899	16,822	9,805,688	7,580,582	2,225,106
4	900-2699	20,995	32,384,583	26,494,811	5,889,772
5	2,700-8,099	6,834	30,245,100	25,801,311	4,443,789
6	8,100-24,299	1,531	19,455,632	17,152,575	2,303,057
7	24,300-72,899	328	15,658,671	14,183,073	1,475,598
8	72,900-218,699	699	91,503,464	84,794,014	6,709,450
9	≥218,700	430	184,235,599	174,124,745	10,110,854
Total		85,571	385,709,457	351,784,276	33,925,181

[modified from Table S1P (Reichwald et al. 2015)]

4 Discussion

In this thesis, I describe the (i) step-wise assembly process of the genome sequence of the short-lived fish *N. furzeri*, (ii) development of a generally applicable method for repeat detection in WGS data (RepARK) and (iii) repeat analysis of the *N. furzeri* genome. For genome sequencing, three sequencing technology generations were employed so that a variety of sequence data types was available for basic genome assembly and repeat analyses. Additionally, optical and genetic mapping data were incorporated and synteny comparisons performed to improve the genome assembly, thus resulting in a chromosome-scale assembly. Using RepARK and further methods, I determined that (i) the repeat content of the *N. furzeri* genome is ~65%, which is extremely high for a vertebrate species, (ii) repeats comprise 35.5% of the genome assembly, which is an underrepresentation due to shortcomings of current assembly programs, (iii) the unique fraction of the *N. furzeri* genome provided in the assembly is ~90%, and (iv) bursts of repeat transposition occurred at several stages in the evolution of the *N. furzeri* genome with indications of an ongoing TE activity.

4.1 Strategies for the *de novo* Genome Assembly of *N. furzeri*

Sequencing technologies have seen a rapid development over the past decades. Sanger sequencing represents the break-through technology that first facilitated sequencing entire genomes. Initially, the sequencing reactions were manually read-out and transformed into sequences. Only later, the read-out was performed by computer programs that assessed and categorized sequence quality, searched for overlaps between reads and allowed for sequence editing (Staden 1979). Upon technology improvements and successful automation efforts, large consortia were formed to sequence and assemble complex genomes of eukaryotes like the yeast (*Saccharomyces cerevisiae*) (Goffeau et al. 1996), worm (*Caenorhabditis elegans*) (C. elegans Sequencing Consortium 1998), fly (*D. melanogaster*) (Adams et al. 2000) or the human genome (Lander et al. 2001; Venter et al. 2001).

Once 2nd generation sequencing technologies became available, more data were produced in shorter periods and at lower expense, which allowed sequencing complex genomes at reasonable costs. However, the short read lengths and the vast amount of data inherent to 2nd generation technologies led to a shift of the work load from sequencing to the assembly process. In particular, repetitive sequences pose a problem on assembling genomes from short reads in regions where the repeat length exceeds the read length. However, it is crucial to reconstruct the genomic sequence accurately as both contiguity and base accuracy of an assembly have an impact on all downstream analyses (Berlin et al. 2015). For sequencing the *N. furzeri* genome, mainly 2nd generation technologies were employed so that respective inherent challenges had to be addressed by applying an appropriate sequencing and assembly strategy, which will be discussed in the following.

Prior to the start of my work, paired-end and mate-pair genomic DNA libraries of *N. furzeri* were sequenced with 454 and Illumina technologies. The typical features of these data (short reads and

vast amount) call for a DBG assembly strategy. Accordingly, the DBG programs *phusion2*¹⁸ and *CLC* were initially tested on male and female *N. furzeri* sequences, which unfortunately resulted in highly fragmented assemblies with up to one million contigs.

To reduce the impact of heterogeneity on the process of genome assembly, it is advisable to sequence inbred, i.e. nearly homozygous, individuals. The *N. furzeri* GRZ strain is highly inbred (Reichwald et al. 2009), and the only heterogeneous region is expected at the sex determination locus. Therefore, individuals of this strain were sequenced and reads of the homogametic, female sex (Valenzano et al. 2009) were used for the initial genome assembly. Building a female-only assembly can also be advantageous for downstream analyses where the focus lies on detecting sex-related sequence differences. To identify the sex determination locus, one can map reads obtained from males onto the female assembly. Positions that show higher than genome-wide average rates of mismatches between male reads and the female reference sequence indicate a sex-related difference (Reichwald et al. 2015).

At the time when initial assembly methods were tested, ALLPATHS-LG was published and showed tremendous improvements over previous programs when performing a *de novo* assembly of the human genome from Illumina WGS data (Gnerre et al. 2011). The program had implemented several novel features. First, a multi-pass correction removes as many as possible sequencing errors from input reads. Second, the program requires overlapping paired-end reads from short DNA fragments, which are then computationally connected to improve the base accuracy in the overlap region and increase the length of the input sequences from ~100 nt to ~170 nt. These accurate fragment sequences serve as substrate to infer long (96 nt) k-mers for building a DBG-like assembly graph. Longer k-mers are advantageous as they decrease the probability to be derived from multiple regions of the genome and thus reduce complexity of the assembly graph (Butler et al. 2008). Last, high coverage of mate-pair data showing different insert lengths are required to facilitate scaffolding.

To meet the requirements of ALLPATHS-LG for both sequence data and computational resources, in total 83-fold paired-end fragment reads and 77-fold mate-pair reads were sequenced, and high-performance computer hardware was purchased. During the process of obtaining the sufficient genome coverage, I carried out test runs of ALLPATHS-LG to evaluate its performance for *N. furzeri* and to determine hardware requirements. For example, a first pass with 50-fold paired-end and 32-fold mate-pair coverage resulted in more scaffolds with a smaller total assembly size compared to the final ALLPATHS-LG run (50-fold paired-end and 50-fold mate-pair coverage, referred to as assembly A). In particular, 40% less scaffolds were built in assembly A but the total assembly length and the scaffold N50 length increased by 4.5% and more than 50%, respectively, illustrating the importance of

¹⁸ <http://sourceforge.net/projects/phusion2/>

providing a sufficient amount of sequence data. Improvements of the ALLPATHS-LG software can also account for better metrics as the program was under constant development at this time.

When preparing libraries and generating sequences that are required as input data for the ALLPATHS-LG, remarkable differences in yield and quality between different fragment types were observed. The preparation of short fragment paired-end Illumina libraries is highly efficient, and sequencing of only two flow cells (i.e. two runs) was sufficient to yield a 83-fold genome coverage. In contrast, mate-pair libraries of 3 kb fragments were difficult to prepare due to several rate-limiting steps. For example, a considerably lower DNA output of the individual mate-pair libraries led to lower sequencing coverage. Therefore, ten mate-pair libraries were prepared and sequenced to obtain a 77-fold genome coverage. Additionally, high numbers of duplicons (identical read pairs) were observed in individual mate-pair datasets. These identical read pairs are supposed to result from sequencing the same DNA fragment and do not add new information to a read set. Therefore, abundant duplicons were removed, which resulted in an overall genome coverage of nearly 50-fold of all mate-pair libraries.

Functionality of ALLPATHS-LGs is generally restricted to Illumina short reads and includes a scaffolding step. However, attempts of the developers were made to use also PacBio data as input, which is currently applicable for bacterial genomes (~ 5 Mb) (Ribeiro et al. 2012). To improve the scaffolding step in the assembly of the large genome of *N. furzeri*, long (8 and 20kb) fragment mate-pairs sequenced by 454 were integrated using the in-house developed program KILAPE. In this process, scaffolds and contigs were connected if multiple mate-pairs support a join. This procedure is often used in scaffolding programs like for example in the popular tool SSPACE (Boetzer et al. 2011) but in addition to that, KILAPE is also repeat aware, in that repetitive regions within reads are excluded during mapping, which prevents false joins of scaffolds. The 454 8 kb and 20 kb mate-pair data reduced the number of scaffolds by more than 50%.

Over the last years, assemblies of many genomes were created by using ALLPATHS-LG and 2nd generation sequencing data, including at least 30 genomes by the developers themselves¹⁹. All are of complex eukaryotic species and have usually an N50 length between 1 Mb and 6 Mb. Among these are a number of fish genomes, for example the spotted gar²⁰, the coelacanth (Amemiya et al. 2013), the northern pike (Rondeau et al. 2014) as well as five African cichlid species (Brawand et al. 2014). Other commonly used assemblers like Velvet, SGA (Simpson and Durbin 2012) or SOAPdenovo usually result in more fragmented assemblies (Salzberg et al. 2012). For the *N. furzeri* genome, an independent, alternative assembly of the GRZ strain was built by merging two *de novo* assemblies of SGA and SOAPdenovo by Valenzano et al. (46,729 scaffolds, 1.02 Gb, N50 247 kb) (Valenzano et al. 2015). Regarding its metrics, it ranks between assemblies A and B presented in this work.

¹⁹ <https://www.broadinstitute.org/software/allpaths-lg/blog/?cat=5>

²⁰ <http://www.ncbi.nlm.nih.gov/nuccore/AHAT000000000>

Of the five steps applied for assembling the *N. furzeri* genome, optical mapping improved assembly contiguity most remarkably as is reflected by a N50 length of 15.8 Mb. In the goat genome project that represents the first complex genome project for which optical mapping was utilized, the N50 is very similar, i.e. 16.3 Mb (Dong et al. 2013). Reaching these N50 lengths is only possible because of a paradigm shift in respect to contiguity of input data - from short-distance sequence patterns to long-distance restriction fragment patterns. For *N. furzeri*, several restriction enzymes were tested by OpGen to select the one (BamHI) that produces an appropriately spaced and sufficient number of restriction fragments. Similar to classical sequence assemblies where overlaps of bases suggest a connection of reads, in optical mapping restriction fragment length patterns are compared and connected if they match. A *de novo* genome assembly-based only on optical maps would be computationally even more complex (Jing et al. 1998) than a *de novo* genome assembly-based on sequences. This is due to a larger space of possibilities for each positional element (the fragment length as rational number vs. “A”, “C”, “G”, “T”), length variations due to inaccurate detection of DNA fragments and uncertainties from false-negative and false-positive restrictions. Therefore, OpGen required long, sequencing-based contigs, which were used as seeds to initiate the map creation that was then followed by an iterative aligning-elongation approach. Providing such long seed contigs may not easily be possible, which can be seen as a drawback of the OpGen method. For the *N. furzeri* assembly, long scaffolds with <5% of Ns from assembly B were selected, *in silico* digested, and used as seeds for the iterative map elongation. OpGen provided maptigs that were elongated at least eight times. The number of iterations is a compromise between too few iterations, where valuable information remain hidden, and to many iterations, which may introduce uncertainties by possible misassemblies.

The optical map (975 Mb) represents only 65% of the estimated genome size (1.5 Gb), for which there are three reasons: (i) The large centromeric and peri-centromeric regions are not expected to be included in the genome assembly. These regions are comprised of two prominent tandem repeats with a respective unit length of 77 bp and 348 bp (Reichwald et al. 2009) that do not contain a BamHI restriction site and remain undetected during optical analysis of the DNA molecules. Even if another restriction enzyme were used that can successfully digest these repeats, the length of their repeat unit is below the resolution threshold (~5 kb fragment length). (ii) Only OpGen scaffolds larger than 250 kb were used as seeds for creating maptigs. Therefore it is possible that maptigs corresponding to regions not represented in the sequence assembly are lacking. (iii) Even though it was intended to create maptigs from a large number of seeds (1,068) some maptigs may either not have made it to the 8th iteration of the maptig assembly or may be below the 30-fold average coverage threshold.

At the time optical mapping was performed for *N. furzeri*, the service was commercially available only from OpGen. In the meantime, another company named BioNano provides a similar service called “Genome Mapping”. The technology was recently used to build an alternative human

genome assembly showing a scaffold N50 lengths of ~30 Mb (Pendleton et al. 2015). A genome map was *de novo* assembled from 50-fold coverage single molecule maps (comparable to OpGen's SMRMs), which served as a backbone to anchor contigs assembled from PacBio data. In contrast to OpGen, BioNano provides the user better access to software (Shelton et al. 2015), and additional programs were developed by other researchers for that kind of data (for example OMBlast²¹) thus opening up the possibility to independently verify or re-assemble optical maps. In general, optical mapping has proven to represent a powerful tool to increase substantially the contiguity of sequence-based genome assemblies in bacteria (Nagarajan et al. 2008), plants (Shearer et al. 2014) and animals (Dong et al. 2013). Moreover, it represents an independent approach to verify the correctness of sequence assemblies (Kawahara et al. 2013). For the *N. furzeri* assembly, a consistency of 87% between sequence scaffolds of assembly B and the OpGen scaffolds was observed (B. Downie, personal communication). One quarter of the inconsistencies was due to low quality of scaffold-mapping alignments whereas the remaining inconsistencies were identified as misassemblies, which were resolved by either truncating the respective scaffolds or omitting them from incorporation into the superscaffolds of assembly C. This quality assessment strongly increased the reliability of the *N. furzeri* genome sequence.

Also incorporating information from genetic maps was an important step to improve the *N. furzeri* genome assembly as it allowed anchoring of scaffolds and superscaffolds to chromosomal units. Different genetic linkage maps of varying size, resolution and marker types were available for the *N. furzeri* genome. The map by Kirschner et al. (Kirschner et al. 2012) was chosen as a basis for genetic map integration because it has the highest number and density of markers (distance of 5.5 cM) and a low number of LGs (n=22) which is close to the number of chromosomes (n=19). The general workflow applied in this work was to order and orient the scaffolds based on their marker sequence according to their positions within the LGs. If scaffolds were short, or if only few markers were present, sometimes no clear-cut decision was possible. Comparing those cases to additional linkage maps either supported an anchoring or not. Accordingly, nearly three quarters of all superscaffolds were successfully anchored to LGs. Moreover, the unambiguous positioning of genetic markers on long superscaffolds supported the merging of three pairs of LGs (i.e. six LGs) to yield a final number of 19 GSCs reflecting the 19 *N. furzeri* chromosomes. Chromosome end annotations provided by OpGen for 24 mapping/superscaffold ends further confirmed the anchoring of many superscaffolds. There was only one case, where a chromosome end annotation would have been placed in the centre of a GSC. This conflicting annotation was discarded because i) the SMRM coverage was below the threshold for a confident annotation and ii) independent FISH analysis showed that neighboring superscaffolds are located on one (the same) chromosome.

²¹ <http://www.hkbic.cuhk.edu.hk/software/omblast>

The anchoring of sequences to GSCs based on genetic maps was done manually in the *N. furzeri* assembly, since there was no appropriate automated software available. Only recently the program ALLMAPS was published that can anchor sequences based on multiple maps of different types and origins, which can also be ranked by priority of particular maps (Tang et al. 2015). This program is to be considered for possible future efforts aimed to improve the genome assembly. Such an analysis could also evaluate marker distances by comparing them between the scaffold level and the linkage group level as the current comparison only considers the marker position but not inter-marker distances.

Genetic maps have been powerful resources in genetics since decades and have supported greatly many genome sequencing projects (Fierst 2015). In a number of efforts to assemble mammalian genomes, more than 90% of the sequences were assigned to chromosomes by using linkage and different physical maps that often contain many thousands of markers (Lewin et al. 2009). Also fish genome assemblies were improved by genetic mapping. Of the compact genome of the pufferfish *Takifugu rubripes* (400 Mb assembly size), 86% was anchored to chromosomes using 1,220 microsatellite markers (Kai et al. 2011). In rainbow trout (1.9 Gb), 54% of the assembly was anchored to chromosomes based on different genetic and physical maps (Berthelot et al. 2014). In medaka (700 Mb), 90% of the nucleotides were anchored to chromosomes using 2,401 SNP markers (Kasahara et al. 2007). For the zebrafish (1.4 Gb), only recently a high-density genetic map with 140,306 SNP markers was created and served as a backbone to anchor 96% to the chromosomes. For *N. furzeri* (950.8 Mb), 77% of the assembly was anchored to 19 GSCs, which is a remarkably high proportion regarding that only 387 markers were used. The incorporation of optical mapping data in the assembly largely accounts for this success as many long range connections were established for the superscaffolds. In general, one can conclude that the more complex these fish genomes are, the more important it is to perform map integration for facilitating genome assemblies of good quality. This is also true for other complex and large genomes like those of grasses. For example, the 5 Gb barley genome required a multitude of physical and genetic maps to assemble its six chromosomes (International Barley Genome Sequencing Consortium 2012).

In recent years, new map construction strategies were developed that use SNPs present in sequence tags adjacent to restriction enzyme cut sites (restriction site associated DNA (RAD) markers). These genomic regions are captured and sequenced in parallel, so that a high number of polymorphic markers can be rapidly obtained to construct a genetic map (Catchen et al. 2011). Using such a map comprised of 16,114 RAD markers, 90% of the platyfish genome assembly (730 Mb) was anchored to chromosomes (Schartl et al. 2013; Amores et al. 2014). Very recently, a genetic map of *N. furzeri* was constructed from 8,399 RAD markers that allowed assigning 35% (2,800 scaffolds, 380 Mb) of an alternative *N. furzeri* assembly (1.02 Gb) to 19 LGs (Valenzano et al. 2015). This RAD map can be used to evaluate the *N. furzeri* SGRs presented in this thesis and to potentially further

improve the genome assembly by anchoring yet unassigned contigs, scaffolds or superscaffolds to chromosomes.

As one quarter of the superscaffolds remained un-anchored after genetic map integration, pairwise synteny comparisons between the *N. furzeri* genome assembly and those of medaka and stickleback were performed. Prior to that, medaka and stickleback were compared to evaluate the applicability of synteny-based scaffolding between fish genomes. The results showed a good overall degree of synteny between those two species (A. Petzold, personal communication), thus suggesting that synteny can be applied for further scaffolding the *N. furzeri* data. Using synteny analysis, all but three of the remaining superscaffolds were connected to the existing GSCs forming 19 SGRs. A successful, synteny-based scaffold placement is highly dependent on the quality of gene annotation of the respective genome assembly. For *N. furzeri*, a transcript catalogue was available (Petzold et al. 2013) and a comprehensive gene annotation that includes 26,141 protein-coding genes was carried out (Reichwald et al. 2015). The synteny-based scaffold placements were validated by comparison with optical mapping data. In nine out of ten randomly selected cases, the suggested placement did not conflict with the maptigs. In one case, a scaffold was larger than a gap in the maptig, which suggests a misassembly of either the maptig or the scaffold (data not shown). To resolve such conflicts, RAD maps may be utilized.

In addition to methods applied and discussed for scaffolding in this work, there are other approaches shown to be efficient for ordering and improving genome assemblies. These make use of long-range information from either so called sister chromatid exchange obtained by single cell sequencing (Hills et al. 2013) or from chromatin interaction obtained by Hi-C data (Lieberman-Aiden et al. 2009). In the latter method, covalent links of neighboring DNA chromatin segments are induced and these linked pairs of DNA are sequenced using paired-ends where one end of the read pair corresponds to the first DNA segment and the second read corresponds to the second segment. Under the assumption that intra-chromosomal links are much more likely than those between two chromosomes, contigs of (classical) sequence assemblies can be grouped by chromosome and, when the Hi-C read pairs are mapped onto the contigs, even ordered (Burton et al. 2013). These methods can also be useful for evaluating the contiguity of genome assemblies (Marie-Nelly et al. 2014). However, Hi-C is based on a sophisticated wet-lab protocol (de Wit and de Laat 2012) and so far has not been widely adopted for genome assembly although it was used to construct genome sequences for humans and the American alligator (Putnam et al. 2015) as well as for *Arabidopsis thaliana* (Xie et al. 2015).

One of the biggest challenges in *de novo* assembling complex genomes is the presence of repeats (Treangen and Salzberg 2012). Usually, 2nd generation reads are shorter than repeat motifs or repeats, which results in fragmented assemblies with many gaps. Moreover, due to high sequence similarity within repeat families, repeats can cause incorrect assemblies because assemblers may make wrong assumptions to connect contigs (Phillippy et al. 2008). Already in the initial characterization of

the *N. furzeri* genome, an exceptionally high repeat content was reported (Reichwald et al. 2009); accordingly, these repeat related challenges were considered and tackled by different means during the genome assembly process presented in this work. In the basic assembly A, ALLPATHS-LG connects overlapping reads with a length of 100 nt to build longer fragments that span short repeats and solve as basis for the assembly graph construction (Gnerre et al. 2011). This strategy allows the usage of 96-mers, which are longer than those applied by other DBG assembly programs and result in a more compact and less ambiguous assembly graph. When working with such long k-mers, a high amount of memory is required.

A second method to address the challenges introduced by repetitive sequences is to use mate-pair reads that span several kb and allow connecting contigs/scaffolds flanked by repeats. Not only the initial ALLPATHS-LG assembly but also following assembly steps benefitted from those long-range libraries. Generally, it is suggested to make use of a broad range of insert sizes in sequencing libraries and also to consider the average repeat length of a particular genome (Wetzel et al. 2011). For *N. furzeri* the most commonly used protocols were applied to produce 3 kb, 8 kb or 20 kb mate-pair libraries. Additionally 108,994 genomic BAC insert ends were sequenced and included for scaffolding. However, their usefulness for scaffolding was limited due to their relative low coverage (0.04-fold). In general, the outlined strategies improve assembly contiguity but often do not add actual sequence information. In the *N. furzeri* genome assembly, it was attempted to fill sequence gaps by using 2nd and 3rd generation data. For example, in assembly step B, 454 reads were used to fill gaps during scaffolding with KILAPE followed by dedicated gap filling with Illumina data, which together resolved 40 Mb of Ns. Additionally after the final genome assembly, a pilot experiment was performed using raw WGS PacBio subreads, which filled 16.7 Mb of gaps that were found to be predominantly composed of repeats. This strongly supports the general assumption that repeats lead to fragmented assemblies (Alkan et al. 2011) and calls for further WGS PacBio sequencing for improving the *N. furzeri* genome assembly.

Repeats cannot only hamper *de novo* assembly but also down-stream analyses based on that assembly. For example, the detection and annotation of genes is affected, because unmasked repeats can lead to false-positive gene annotations (Yandell and Ence 2012). Many genome-wide analyses are based on mapping of sequenced reads. Finding (nearly) identical hits of single reads at multiple locations in a genome challenges the mapping algorithm and might lead to wrong assignments (Reinert et al. 2015). In turn, wrong mappings can increase the number of false positives in variation detection or genotyping (Treangen and Salzberg 2012). Evaluating these problems is out of the scope of this work but needs careful consideration in every genome project.

4.2 Repeat Identification and the Development of RepARK

Repeat identification and genome assembly are two tightly connected fields. Because unassembled 1st and 2nd generation reads are often too short to span entire repeat copies, the identification of repeats is

normally based on assembled sequences. This approach is usually advantageous, since the probability for capturing full-length repeats is higher in assembled data. Also, structural properties can be deduced more easily and this may assist in classifying repetitive elements. However, most of the currently available assemblies of complex genomes are not in a finished state, meaning that a considerable fraction of the genomic sequence may be lacking. These missing sequences are mainly copies of repeats (Tang 2007). This is a vicious circle situation, which one can only escape by uncoupling repeat identification from genome assemblies. The unassembled raw data of sequencing experiments is supposed to represent best the content of any given genome. Therefore, approaches were developed that detect and assemble repeats from sequencing reads but are either designed for 1st generation sequencing data (ReAS) or limited to certain software frameworks, e.g. the Galaxy framework (RepeatExplorer). The recently published program Transposome is based on the graph concept of RepeatExplorer while it is faster and more flexible in terms of its running environment. It detected 17 of the 20 major TE families in the highly repetitive maize genome using a genome coverage of <1-fold (Staton and Burke 2015). By contrast, the RepARK pipeline described in this thesis generates repeat libraries based on the method of k-mer counting, which is used in several fields of sequence analysis (Pevzner et al. 2001). Like Transposome, RepARK is a stand-alone pipeline but even more flexible as its components are easily exchangeable. Because Transposome was published after RepARK, the performance of both programs was not yet compared using the same data basis.

The *D. melanogaster* genome assembly was employed to develop the RepARK pipeline and to demonstrate its “usability” because that genome has both a moderate size and repeat content and it was improved several times over more than a decade. The method was validated by using both simulated and real sequence data of *D. melanogaster* and also the application to larger genomes was shown with human data. For *D. melanogaster*, the overall lengths of the four RepARK repeat libraries are greater than that found in RepBase (0.87-4.3 Mb vs. 0.7 Mb), and >90% of consensi in each RepARK library are repetitive. Each library, with the exception of RepeatScout, masked between 22% and 32% of the reference genome, which indicates a generally higher fraction compared to early reassociation kinetics (12%) (Manning et al. 1975) or recent analyses of unassembled data (18%) (Krassovsky and Henikoff 2014). The difference between these two values and the value obtained in the current work can be explained by the fact that the highly repetitive and fragmented heterochromatin sequences (Smith et al. 2007) were also included in the reference sequence which was used for evaluation of the repeat libraries. However, only a small fraction of the reference masked with a RepARK library can be subsequently identified by RepBase as a repeat (0.18-1.18%), indicating that the majority of RepBase repeats in the genome can be identified using the RepARK method.

Regarding the difference between real and simulated data, it was generally found that more consensi were assembled in RepARK libraries using the simulated dataset than using the real dataset.

Such a discrepancy could result from assembly errors in the reference sequence leading to an artificial variability of certain motifs. This may result in particular from including the U and Uextra chromosomes for read simulation, as they are hotspots for assembly errors (Smith et al. 2007). Alternatively, real sequencing data are subjected to various technological biases leading to the underrepresentation of particular motifs (e.g. GC-rich or heterochromatin sequence, which are both regions of high and low repeat content (Dohm et al. 2008)). Finally, it is possible that this discrepancy is due to actual genomic differences between the reference and the DNA sample sequenced such as single nucleotide variations, copy number variations or segmental duplications.

Focusing on the `wgs-assembler`, a difference between simulated and real data is observed as it produced a comprehensive repeat library in almost all metrics when using simulated data. With real data, the assembler generated the longest library amongst all real data libraries but performed much worse, in particular when masking the reference genome (22% vs. 27-32% for the other non-RepeatScout libraries). This suggests a low sensitivity in repeat identification. It is also important to note that the RepeatScout-based method, which is the most popular state-of-the-art approach for *de novo* generation of repeat libraries, was the least effective at generating comprehensive repeat libraries of all methods examined. The fact that RepeatScout identified only a small amount of repeats in the Velvet-based *de novo* genome assemblies underscores the dependence of RepeatScout or similar detection methods on a high quality reference assembly which, for complex genomes, is difficult to obtain using only 2nd generation sequence data.

Although 26-35% of the RepBase consensi showed a completeness of <50% in the RepARK libraries, one third of these consensi belonged to the RepBase group “remaining” (Figure 11), which contains for example consensi of bacterial transposons (Broom et al. 1995). During the process of building DmRepBase, those bacterial elements were extracted from RepBase by default as they are formally allocated to all available species including *D. melanogaster*. This precaution is motivated by possible contaminations during sequencing protocols which involve bacterial cloning. Because cloning-free 2nd generation sequencing was used for RepARK such artifact sequences are not expected in the data. Moreover, the DmRepBase library contains ancestral repeat consensi, which may not be repetitive or represented at all in the reference genome and therefore were not detected as repeats by RepARK. Finally, it is also possible that during the process of repeat consensus assembly, highly divergent repeat motifs may cause excessive fragmentation of the assembly graph resulting in short sequences which are either not reported by the assembler or fall below the length cut-off of 50 bp.

More genome sequence is masked by RepeatMasker using the RepARK libraries than using DmRepBase library (1.6-4.5% additional sequence). Part of this additional masked sequence can be explained by the observation that RepARK consensi also includes SDs. In contrast, RepBase libraries contain only simple and genome-wide dispersed repeats. Usually, SDs are detected using traditional

whole-genome alignment methods based on criteria ($>90\%$ identity, >1 kb) which exclude shorter and more divergent sequences. This could explain some of the putative novel SD events identified using the RepARK libraries. Additionally, the use of whole-genome alignments to detect SDs runs the risk of false negatives due to assembly errors in the reference sequence. Holding SDs in the repeat library can have an influence on subsequent gene annotation as it may be possible that genes located in duplicated regions are masked and therefore are not detected. These consequences should be kept in mind for downstream analyses when including SDs for repeat masking. However, given the high ratio of fully mappable consensi, this data further underpins the conclusion that the consensi produced by RepARK are both highly specific and sensitive for detection of repetitive elements of a given genome.

The bias toward DNA transposon classification by TEclass for the RepARK and wgs-assembler libraries represents a limitation for accurately annotating repeat classes in a genome. This behavior is most likely due to the highly fragmented nature of such libraries, which may present a challenge for some of the annotation models implemented in TEclass. Revising these models may produce more accurate classification of highly fragmented repeat libraries such like those investigated here. Alternatively, construction of longer repeat consensi (such as those found in RepARK library generated by Velvet) or the restriction of TEclass to longer consensi (>100 bp) can also improve repeat classification. Regardless to further improvements, precise examination of repeat evolution in newly assembled genomes requires closer, manual examination. Nevertheless, the consensi of RepARK libraries can be used to identify and isolate repetitive genomic elements with high accuracy and to provide a first pass genome annotation.

When applying RepARK to experimentally-derived human reads, a similar rate of true positives is observed as for *D. melanogaster*. This shows that the method is also applicable to more complex vertebrate genomes. Unexpectedly, the entire EBV genome was found within the RepARK library. This can be explained by the fact that EBV was used to establish the cell line from which the human DNA was isolated and sequenced. Additionally, by re-mapping the k-mers onto the EBV sequences, the copy number of the virus genome was estimated to $n=17$. A recent study found three EBV integrations into the genome of that particular sample (Mak et al. 2016). The difference of 14 copies may result from additional EBV genomes as they usually exist as a circular episome within the nucleus (Morissette and Flamand 2010). These findings suggest that RepARK may also represent a novel method to identify multi-copy contaminants within a DNA dataset and may find future application not only as a repeat library generator, but also as a diagnostic tool.

4.3 Repeat Content of the *N. furzeri* Genome

The genome of *N. furzeri* shows a remarkable signature of repeats (Reichwald et al. 2009). Even in the initially analyzed, relatively small sequence sample, 45% was identified to be repetitive with an extraordinary high TR fraction of 21%. Obviously, these observations had to be further investigated by

using a broader range of datasets and by different means. For this, data of three generations of sequencing technologies and data that was processed to different stages (i.e. raw reads and assembled sequences) allowed a comprehensive repeat analysis. In cases like *N. furzeri* where the distance to most closely related known genome (medaka) is rather large, it is particularly important to thoroughly investigate the repeat fraction by different means. It is not sufficient to only rely on homology comparison to repeat databases of other species. Thus, *de novo* repeat detection methods are needed to avoid missing the species-specific elements (Platt et al. 2016).

I created three independent *N. furzeri* repeat libraries using different datasets and compared their potential to detect repeats in the genome assembly. To create the library RModLib, RepeatModeler performed a *de novo* detection of DRs within the genome assembly and automatically classified these sequences by comparison to RepBase repeats. Library construction based on the Sanger sequences (SangerLib) was done in two steps where known RepBase repeats as well *de novo* RepeatScout consensi were collected. Therefore, the initial SangerLib contained a huge fraction of already classified RepBase repeats that show similarities to *N. furzeri*. This result however probably does not reflect the true situation in *N. furzeri* as the detected repeats are only copies from RepBase coming from different species. Thus, those were excluded and only *de novo* consensi from RepeatScout were considered for further analyses. The RepARKLib did not contain any initial classification. Each library was separately filtered for short sequences and redundancy, which dramatically decreased the size of the RepARKLib by the factor of 3.5 and the number of consensi by the factor of 8.4, but did not affect the two other libraries that drastically. This reduction of complexity of the RepARKLib is also important for practical reasons, as this decreases the runtime of RepeatMasker using that library.

The classification of the repeat consensi was carried out as an iterative process on the CombinedLib. Classified consensi were removed from the input of a subsequent round. In contrast to a strategy which constantly uses all consensi from the CombinedLib as input of each classification step, the used approach prevents conflicting results. However, this also rules out possible confidence enrichments when different classifications steps would agree on the same result. Nevertheless only high quality classifications of the consensi were accepted while uncertain ones were re-submitted to the subsequent rounds again. The different steps of classification start with a broader spectrum of subjects for comparison and end with a manual investigation where only selected families are examined. This narrowing is also reflected by the decreasing fraction of classified consensi after each step. In total, one third of all CombinedLib consensi was classified, of which Censor+RepBase contributed most consensi (4,509; 18%) followed by Censor+FishRepLib (2,228; 9%) and the manual classification (230; 1%). The fact that two thirds of the CombinedLib remains unknown seems dissatisfying but is consistent with results from the publication of the classification tool REPCLASS, which classified 24-33% of repeat consensi from three nematode species. When REPCLASS was

applied to different *Drosophila* genomes, it classified 50-57% of the consensi. The authors explain this difference with the varying grade of repeat knowledge, which is much more advanced for the *Drosophila* species compared to nematode genomes (Feschotte et al. 2009). Also for *Nothobranchius* and related species, no specific repeat annotation is currently available in RepBase. Consequently, the unclassified consensi might represent novel *Nothobranchius* lineage-specific repeats which further underscores the diversity of repetitive elements among species.

After having applied REPCLASS to an earlier version of the SangerLib (Koch 2010), in the current work, I omitted its usage for the classification of the CombinedLib for three reasons: (i) The previous analysis in the Sanger data successfully classified only 13% of the 3,386 unclassified consensi which is around half of the fraction expected from the literature (Feschotte et al. 2009). (ii) REPCLASS demands for a reference sequence to which it aligns all repeat consensi to detect possible flanking structural repeat features. However, this interconnection with the genome assembly was not intended as it was desired to classify the CombinedLib as independent from a reference as possible. (iii) Mainly due the reference alignments and the evaluation of these results, the runtime of REPCLASS was relatively high for the 120 Mb Sanger sample and was expected to be much higher for the entire *N. furzeri* genome assembly. A possible future usage of REPCLASS is conceivable for a fraction of the CombinedLib or single repeat consensi that might be of further interest.

A comprehensive repeat composition analysis of the *N. furzeri* genome assembly was conducted in a two-step pipeline. Because TRs differ in structure and distribution, they were separately analyzed prior to the DR detection with the CombinedLib. A fraction of two percent of the total genome assembly is composed of TRs which is in great contrast to >20% in the initial characterization of the *N. furzeri* genome (Reichwald et al. 2009) and also to the 15-28% observed from raw sequencing data (Figure 19). This difference can be explained by the fact that the first analysis was based on longer Sanger sequences whereas the current genome assembly is predominantly built from shorter Illumina reads that were assembled by a DBG assembler which tends to collapse short repetitive motifs. It is therefore highly likely that only a small fraction of TRs was assembled correctly. In the initial characterization of the genome, two prominent TRs were described (a G+C rich 77 bp minisatellite and a G+C rich 348 bp satellite) which made up large parts of the total TR fraction and which localized preferentially to the peri-centromeric regions (Reichwald et al. 2009). Both repeats are present in the different sequencing datasets (Figure 19). However, no substantial numbers of these TRs were found in the genome assembly, most likely because they were not assembled due to limitations of the assembly program mentioned earlier in combination with the short reads.

The second step of the repeat detection pipeline found 33.4% DRs in the genome assembly. This is more than the 25% reported in the initial 5.4 Mb WGS fraction (Reichwald et al. 2009) but less than the 39-49% observed in the different raw datasets (Figure 19) which can be explained by either the limited size of the 5.4 Mb sample or the difficulties during *N. furzeri* genome assembly (Treangen

and Salzberg 2012). Nearly half of the DRs in the genome assembly remained unclassified which is comparable to novel repeat fraction of the medaka repeat content (Kasahara et al. 2007), suggesting that several TE families are yet uncharacterized in teleost fish genomes. Among the classified *N. furzeri* repeat consensi, LINE elements were most prominent (8.4%), followed by DNA (5.8%), LTR (1.9%) and SINE (1%). This class distribution is comparable to that of fugu or stickleback where LINE elements are the dominant class and contrary to many other teleost genomes where more DNA transposons than LINEs or LTRs are observed. A high proportion of LINE elements is also present in mammalian genomes which in contrast lack DNA transposons (Chalopin et al. 2015). When the number of TE superfamilies is considered, its range and diversity is as broad as in other teleost genomes (Volf et al. 2003) while in mammalian and amphibian genomes, a smaller TE diversity is seen (Chalopin et al. 2015). This difference can be explained for compact genomes (e.g. birds) by the general evolutionary trend of genome size reduction (Andrews et al. 2009) or in large genomes by the competition of active TEs fighting for genomic resources (Abrusan and Krambeck 2006).

Although a large fraction of repeat consensi remained unknown, those that were assigned to a class are valuable helpers in understanding the evolutionary history of TEs. For this, the classified consensi are aligned to the genome assembly and the difference between the members of a family is determined. Since the vast majority of the repeats are no longer active in transposition, and have no other obvious function, they will accumulate mutations at the neutral rate. Thus, sequences of more recently transposing members are more similar to their source sequence than those of elements transposed earlier. The results of such analyses give profiles specific for particular genomes and allow a comparison of their TE evolution (Chalopin et al. 2015). For example, in mammalian genomes, a recent and possibly still ongoing proliferation of SINE, LINE or LTR retrotransposons is observed which is represented by a relatively evenly elevated profile without distinct peaks. In contrast, many other vertebrate genomes including teleosts have clear peaks indicating distinct bursts of transposition interleaved by periods with low TE activity. Of those, the zebrafish genome shows a single remarkable and very narrow peak of DNA transposons while the profiles of medaka, cod or tilapia have two peaks. Although the class composition is different, the analysis of *N. furzeri* also shows two major peaks, one of LINE retrotransposons and another of DNA transposons. The very recent expansion of LINE elements suggests an ongoing TE activity in the *N. furzeri* genome while the burst of DNA elements seems to date back to an earlier event in evolution. A similar ongoing proliferation of LINE L1 elements is observed in the mouse genome (Goodier et al. 2001). Other mammalian genomes show comparable L1 activity although these elements are not as highly proliferative as in mouse (Deininger et al. 2003). With a fraction of nearly 5% of the genome assembly, LINE L2 is the most abundant superfamily in *N. furzeri* while they are also detected in other fish genomes like rainbow trout (1.3%) (Berthelot et al. 2014) or Nile tilapia (1.9%) (Brawand et al. 2014) and are always more abundant than L1 elements in these genomes. In the coelacanth genome, L2 elements occupy a relatively small

genomic fraction of 1.3% (Chalopin et al. 2014) but it was shown that they are still transcriptionally active in multiple tissues (Forconi et al. 2014).

Taken the TR and DR content together, 35.5% of the assembly was identified to be repetitive. In contrast, all analyzed raw datasets suggested that the genome-wide repeat content is clearly higher, ranging from 56% to 70% (Figure 19). This would lead one to assume that these repeats are either hidden in the N-stretches of the assembly or are entirely absent from the assembly. As indicated earlier, different gap filling attempts resolved more than 50 Mb. I analyzed in particular the proportion of gaps filled by PacBio data and found that it was predominantly composed of repeats (83.3%). In a gap filling attempt of the African cichlid *Metriaclimbra zebra* (1 Gb genome, 849 Mb ALLPATHS-LG assembly, 16.5-fold PacBio subread genome coverage), 90 Mb of gaps were resolved. For these gaps, a repeat content of 70% has been calculated (Conte and Kocher 2015) which further encourages PacBio sequencing efforts for *N. furzeri*.

The PacBio filled gaps confirm the WGS-based estimate of repeat richness of the *N. furzeri* genome and served also for an assessment of assembly completeness. When they were ordered by length, longer gaps showed a larger fraction of repeats. A logarithmic function, which was fitted to the repeat content of the size ordered gaps, suggested that missing regions >72.9 kb are almost entirely composed of repeats (>92%). This in turn allowed to conclude that 90% of the unique and ~40% of repetitive sequences are contained in the current genome assembly (chapter 3.3.6).

The size of the *N. furzeri* genome was estimated to 1.5 Gb (Reichwald et al. 2009) while the genome assembly contains 1.24 Gb. Even for large international genome projects like those for human or mouse, it is currently still not feasible with reasonable means to assemble a genome of this size and with a respective repeat content in its whole entirety (Treangen and Salzberg 2012). Often, a rough rating of assembly progress or quality is communicated with terms like “draft” and “finished”. Even when an assembly is called finished this term usually refers to the euchromatic portion of the genome. When the “finished” human genome was published in 2004, still 341 gaps remained of which heterochromatin had an portion of only 28 Mb while heterochromatin gaps were estimated to 198 Mb (International Human Genome Sequencing Consortium 2004; Lander 2011). A more precise classification was suggested by members of sequencing institutes and consortia to also appreciate assemblies that are better than draft or ones that are actually not really finished (Chain et al. 2009). The authors formulated six stages starting with (i) “Standard Draft” as the minimum quality that is allowed to submit assemblies to the databases, over (ii) “High Quality Draft” and (iii) “Improved High Quality Draft” without or including (iv) “Annotation-Directed Improvement” requiring additional manual efforts to fill gaps, correct errors or perform genome annotation, up to the levels of (v) “Noncontiguous Finished” and (vi) “Finished” while the last requires a fully reviewed sequence with all repeats resolved and allows only one error per 100 kb. Based on this classification, the *N. furzeri* assembly can be at least categorized as a high-quality draft because it contains 98% of eukaryotic core

genes, 96% of *N. furzeri* transcript contigs (Reichwald et al. 2015) and harbors 90% of unique sequences (this work). On top of those measures, the assembly is featured by 87% of the assembled bases anchored to 19 chromosomes and different gap filling methods were successfully applied and qualifies it for the status “Improved High-Quality Draft” with “Annotation Directed Improvements” (Chain et al. 2009).

5 Conclusions and Outlook

The results reported in my PhD thesis are of great value for currently ongoing efforts to identify genetic factors and genomic entities that affect the *N. furzeri* lifespan. I contributed to the genome assembly and its analysis. The release of these data was long awaited in the scientific community. Based on the genome sequence as well as the provided annotation of genes and repeats, systematic genome-wide analyses are now possible.

The program RepARK was developed since there was a strong demand for bioinformatics tools adapted to detect repetitive sequences in unassembled 2nd generation sequencing data. RepARK considerably improved the repeat analysis in the genome assembly of *N. furzeri* and since its publication in 2015 has been applied to other genome projects (Fitak et al. 2016).

Although the *N. furzeri* genome assembly presented in this work is of high quality, improvements are already ongoing or planned. With recently released tools (Shelton et al. 2015) that allow an OpGen-independent analysis of optical mapping data, quality assessments are possible which potentially result in improvements of the presented assembly. Further, the recently published high-density RAD tag map (Valenzano et al. 2015) and the outcome of similar high-density maps are to be used for validating the current assembly and for improving its contiguity. In addition, it will be attempted to reduce the number of gaps by performing additional PacBio sequencing and applying software tools like PBJelly to fill in sequencing gap. Each improved assembly will require an update of the gene and repeat annotation possibly resulting in the identification of new exons and genes as well as a more complete repeats.

The data presented in this work suggest that TEs are still active in the *N. furzeri* genome, which is supported by the presence of gene models annotated as repeat derived (Reichwald et al. 2015). To address this, genome-wide systematic investigations are warranted. These could be coupled with TE expression analyses in young and old *N. furzeri* similar to experiments performed in mice where it was shown that older animals show a higher TE activity (De Cecco et al. 2013).

The analysis of global DNA methylation in the *N. furzeri* genome is another interesting topic related to this work and would benefit from the repeat library produced here. Highly abundant repeat elements, for example LINEs, could be analyzed by bisulfite amplicon NGS. This is particular interesting in fish, as they show a generally higher DNA methylation level than mammals (Jabbari et al. 1997). In addition, this and other genome-wide methods can be applied to identify (and potentially manipulate) methylation differences between young and old animals, which is intensively studied in mammalian and human aging (Wilson et al. 1987).

6 References

- Abrusan G, Grundmann N, DeMester L, Makalowski W. 2009. TEclass--a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**(10): 1329-1330.
- Abrusan G, Krambeck HJ. 2006. Competition may determine the diversity of transposable elements. *Theoretical population biology* **70**(3): 364-375.
- Adams MD Celniker SE Holt RA Evans CA Gocayne JD Amanatides PG Scherer SE Li PW Hoskins RA Galle RF et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287**(5461): 2185-2195.
- Albig W, Kioschis P, Poustka A, Meergans K, Doenecke D. 1997. Human histone gene organization: Nonregular arrangement within a large cluster. *Genomics* **40**(2): 314-322.
- Alkan C, Sajjadian S, Eichler EE. 2011. Limitations of next-generation genome sequence assembly. *Nature methods* **8**(1): 61-65.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**(3): 403-410.
- Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, Maccallum I, Braasch I, Manousaki T, Schneider I, Rohner N et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**(7445): 311-316.
- Amores A, Catchen J, Nanda I, Warren W, Walter R, Scharl M, Postlethwait JH. 2014. A RAD-Tag Genetic Map for the Platyfish (*Xiphophorus maculatus*) Reveals Mechanisms of Karyotype Evolution Among Teleost Fish. *Genetics* **197**(2): 625-641.
- Andrews CB, Mackenzie SA, Gregory TR. 2009. Genome size and wing parameters in passerine birds. *Proceedings Biological sciences / The Royal Society* **276**(1654): 55-61.
- Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J. 2014. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature biotechnology*.
- Aston C, Mishra B, Schwartz DC. 1999. Optical mapping and its potential for large-scale sequencing projects. *Trends in Biotechnology* **17**(7): 297-302.
- Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature reviews Genetics* **7**(7): 552-564.
- Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome research* **11**(6): 1005-1017.
- Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research* **12**(8): 1269-1276.
- Bartakova V, Reichard M, Janko K, Polacik M, Blazek R, Reichwald K, Cellerino A, Bryja J. 2013. Strong population genetic structuring in an annual fish, *Nothobranchius furzeri*, suggests multiple savannah refugia in southern Mozambique. *BMC evolutionary biology* **13**: 196.
- Batzoglou S. 2002. ARACHNE: A Whole-Genome Shotgun Assembler. *Genome research* **12**(1): 177-189.

- Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**(2): 573-580.
- Bergman CM, Quesneville H. 2007. Discovering and detecting transposable elements in genome sequences. *Briefings in bioinformatics* **8**(6): 382-392.
- Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology* **33**(6): 623-630.
- Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noel B, Bento P, Da Silva C, Labadie K, Alberti A et al. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature communications* **5**: 3657.
- Boeke JD, Corces VG. 1989. Transcription and reverse transcription of retrotransposons. *Annual Reviews in Microbiology* **43**(1): 403-434.
- Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**(4): 578-579.
- Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome biology* **13**(6): R56.
- Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a Genetic-Linkage Map in Man Using Restriction Fragment Length Polymorphisms. *American journal of human genetics* **32**(3): 314-331.
- Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, Simakov O, Ng AY, Lim ZW, Bezault E et al. 2014. The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513**(7518): 375-381.
- Britten RJ, Kohne DE. 1968. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* **161**(3841): 529-540.
- Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *American journal of human genetics* **63**(3): 861-869.
- Broom JE, Hill DF, Hughes G, Jones WA, McNaughton JC, Stockwell PA, Petersen GB. 1995. Sequence of a transposon identified as Tn1000 (gamma delta). *DNA sequence : the journal of DNA sequencing and mapping* **5**(3): 185-189.
- Burton FH, Loeb DD, Voliva CF, Martin SL, Edgell MH, Hutchison CA. 1986. Conservation throughout mammalia and extensive protein-encoding capacity of the highly repeated DNA long interspersed sequence one. *Journal of Molecular Biology* **187**(2): 291-304.
- Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature biotechnology* **31**(12): 1119-1125.
- Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research* **18**(5): 810-820.
- C. elegans Sequencing Consortium T. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**(5396): 2012-2018.

- Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. 2011. Stacks: building and genotyping Loci de novo from short-read sequences. *G3* **1**(3): 171-182.
- Cellerino A, Valenzano DR, Reichard M. 2015. From the bush to the bench: the annual *Nothobranchius* fishes as a new model system in biology. *Biological reviews of the Cambridge Philosophical Society*.
- Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, Patel S, Adams M, Champe M, Dugan SP, Frise E et al. 2002. Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome biology* **3**(12): RESEARCH0079.
- Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C et al. 2009. Genome project standards in a new era of sequencing. *Science* **326**(5950): 236-237.
- Chaisson MJ, Brinza D, Pevzner PA. 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome research* **19**(2): 336-346.
- Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC bioinformatics* **13**: 238.
- Chalopin D, Fan S, Simakov O, Meyer A, Scharl M, Volff JN. 2014. Evolutionary active transposable elements in the genome of the coelacanth. *Journal of experimental zoology Part B, Molecular and developmental evolution* **322**(6): 322-333.
- Chalopin D, Naville M, Plard F, Galiana D, Volff JN. 2015. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome biology and evolution* **7**(2): 567-580.
- Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods* **10**(6): 563-569.
- Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E, Dewdney J, Reuber TL, Stammers M, Federspiel N et al. 1999. Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. *Nature genetics* **23**(2): 203-207.
- Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**(7167): 203-218.
- Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature nanotechnology* **4**(4): 265-270.
- Compeau PE, Pevzner PA, Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. *Nature biotechnology* **29**(11): 987-991.
- Conte MA, Kocher TD. 2015. An improved genome reference for the African cichlid, *Metriaclicma zebra*. *BMC genomics* **16**(1): 724.
- Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nature reviews Genetics* **10**(10): 691-703.
- Craig NL. 1995. Unity in Transposition Reactions. *Science* **270**(5234): 253-253.
- Crollius HR. 2000. Characterization and Repeat Analysis of the Compact Genome of the Freshwater Pufferfish *Tetraodon nigroviridis*. *Genome research* **10**(7): 939-949.

- De Cecco M, Criscione SW, Peterson AL, Neretti N, Sedivy JM, Kreiling JA. 2013. Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. *Aging* **5**(12): 867-883.
- de la Bastide M, McCombie WR. 2007. Assembling genomic DNA sequences with PHRAP. *Current protocols in bioinformatics / editorial board, Andreas D Baxevanis [et al]* **Chapter 11**: Unit11 14.
- de Wit E, de Laat W. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes & development* **26**(1): 11-24.
- Deininger PL, Moran JV, Batzer MA, Kazazian HH. 2003. Mobile elements and mammalian genome evolution. *Current opinion in genetics & development* **13**(6): 651-658.
- Deininger PL, Schmid CW. 1976. An electron microscope study of the DNA sequence organization of the human genome. *Journal of Molecular Biology* **106**(3): 773-790.
- Dewey CN. 2007. Aligning multiple whole genomes with Mercator and MAVID. *Methods in molecular biology* **395**: 221-236.
- Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**(6570): 152-154.
- Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research* **36**(16): e105.
- Dong Y, Xie M, Jiang Y, Xiao N, Du X, Zhang W, Tosser-Klopp G, Wang J, Yang S, Liang J et al. 2013. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nature biotechnology* **31**(2): 135-141.
- Edgar RC, Myers EW. 2005. PILER: identification and classification of genomic repeats. *Bioinformatics* **21 Suppl 1**: i152-158.
- Ehrlich J, Sankoff D, Nadeau JH. 1997. Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics* **147**(1): 289-296.
- Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**(5910): 133-138.
- Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics* **9**: 18.
- English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS one* **7**(11): e47768.
- Erliandri I, Fu H, Nakano M, Kim JH, Miga KH, Liskovych M, Earnshaw WC, Masumoto H, Kouprina N, Aladjem MI et al. 2015. Replication of alpha-satellite DNA arrays in endogenous human centromeric regions and in human artificial chromosome. *Nucleic acids research* **42**(18): 11502-11516.
- Evgen'ev MB, Arkhipova IR. 2005. Penelope-like elements--a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenetic and genome research* **110**(1-4): 510-521.

- Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D. 2009. Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome biology and evolution* **1**: 205-220.
- Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annual review of genetics* **41**: 331-368.
- Fierst JL. 2015. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Frontiers in genetics* **6**: 220.
- Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends in Genetics* **5**: 103-107.
- Fitak RR, Mohandesan E, Corander J, Burger PA. 2016. The de novo genome assembly and annotation of a female domestic dromedary of North African origin. *Molecular ecology resources* **16**(1): 314-324.
- Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in de novo annotation approaches. *PloS one* **6**(1): e16526.
- Forconi M, Chalopin D, Barucca M, Biscotti MA, De Moro G, Galiana D, Gerdol M, Pallavicini A, Canapa A, Olmo E et al. 2014. Transcriptional activity of transposable elements in coelacanth. *Journal of experimental zoology Part B, Molecular and developmental evolution* **322**(6): 379-389.
- Gao S, Sung WK, Nagarajan N. 2011. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *Journal of computational biology : a journal of computational molecular cell biology* **18**(11): 1681-1691.
- Garey MR, Johnson DS. 1979. Computers and intractability: a guide to the theory of NP-completeness. 1979. San Francisco, LA: Freeman.
- Gebhard W, Meitinger T, Höchtel J, Zachau HG. 1982. A new family of interspersed repetitive DNA sequences in the mouse genome. *Journal of Molecular Biology* **157**(3): 453-471.
- Gemayel R, Vences MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual review of genetics* **44**: 445-477.
- Genade T, Benedetti M, Terzibasi E, Roncaglia P, Valenzano DR, Cattaneo A, Cellerino A. 2005. Annual fishes of the genus *Nothobranchius* as a model system for aging research. *Aging cell* **4**(5): 223-233.
- Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America* **108**(4): 1513-1518.
- Goecks J, Nekrutenko A, Taylor J, Galaxy T. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology* **11**(8): R86.
- Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M et al. 1996. Life with 6000 genes. *Science* **274**(5287): 546, 563-547.
- Goodier JL, Ostertag EM, Du K, Kazazian HH, Jr. 2001. A novel active L1 retrotransposon subfamily in the mouse. *Genome research* **11**(10): 1677-1685.

References

- Graf M, Hartmann N, Reichwald K, Englert C. 2013. Absence of replicative senescence in cultured cells from the short-lived killifish *Nothobranchius furzeri*. *Experimental gerontology* **48**(1): 17-28.
- Green P, Ewing B. 2002. Crossmatch <http://www.phrap.org>.
- Gritsenko AA, Nijkamp JF, Reinders MJ, de Ridder D. 2012. GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies. *Bioinformatics* **28**(11): 1429-1437.
- Hartmann N, Reichwald K, Lechel A, Graf M, Kirschner J, Dorn A, Terzibasi E, Wellner J, Platzer M, Rudolph KL et al. 2009. Telomeres shorten while Tert expression increases during ageing of the short-lived fish *Nothobranchius furzeri*. *Mechanisms of ageing and development* **130**(5): 290-296.
- Hills M, O'Neill K, Falconer E, Brinkman R, Lansdorp PM. 2013. BAIT: Organizing genomes and mapping rearrangements in single cells. *Genome medicine* **5**(9): 82.
- Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994. Fast Folding and Comparison of Rna Secondary Structures. *Monatsh Chem* **125**(2): 167-188.
- Honjo T. 1983. Immunoglobulin genes. *Annual review of immunology* **1**: 499-528.
- Houck CM, Rinehart FP, Schmid CW. 1979. A ubiquitous family of repeated DNA sequences in the human genome. *Journal of Molecular Biology* **132**(3): 289-306.
- Huang X, Madan A. 1999. CAP3: A DNA sequence assembly program. *Genome research* **9**(9): 868-877.
- Hutchison CA, 3rd. 2007. DNA sequencing: bench to bedside and beyond. *Nucleic acids research* **35**(18): 6227-6237.
- International Barley Genome Sequencing Consortium T. 2012. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**(7426): 711-716.
- International Human Genome Sequencing Consortium T. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011): 931-945.
- Jabbari K, Cacciò S, Païs de Barros JP, Desgrès J, Bernardi G. 1997. Evolutionary changes in CpG and methylation levels in the genome of vertebrates. *Gene* **205**(1-2): 109-118.
- Jing J, Reed J, Huang J, Hu X, Clarke V, Edington J, Housman D, Anantharaman TS, Huff EJ, Mishra B et al. 1998. Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America* **95**(14): 8046-8051.
- Jubb R. 1971. A new *Nothobranchius* (Pisces, Cyprinodontidae) from Southeastern Rhodesia. *J Am Killifish Assoc* **8**: 314-321.
- Jurka J. 2000. Repbase Update: a database and an electronic journal of repetitive elements. *Trends in Genetics* **16**(9): 418-420.
- Jurka J, Kapitonov VV, Kohany O, Jurka MV. 2007. Repetitive sequences in complex genomes: structure and evolution. *Annual review of genomics and human genetics* **8**: 241-259.
- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**(1-4): 462-467.
- Jurka J, Klonowski P, Dagman V, Pelton P. 1996. Censor—a program for identification and elimination of repetitive elements from DNA sequences. *Computers & Chemistry* **20**(1): 119-121.

- Kai W, Kikuchi K, Tohari S, Chew AK, Tay A, Fujiwara A, Hosoya S, Suetake H, Naruse K, Brenner S et al. 2011. Integration of the genetic map and genome assembly of fugu facilitates insights into distinct features of genome evolution in teleosts and mammals. *Genome biology and evolution* **3**: 424-442.
- Kapitonov VV, Jurka J. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends in Genetics* **23**(10): 521-529.
- Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**(7145): 714-719.
- Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu JZ, Zhou SG et al. 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6**.
- Kelley DR, Schatz MC, Salzberg SL. 2010. Quake: quality-aware detection and correction of sequencing errors. *Genome biology* **11**(11): R116.
- Kent WJ. 2002. BLAT---The BLAST-Like Alignment Tool. *Genome research* **12**(4): 656-664.
- Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin C-S, Raponi NA, Rank DR, Li J et al. 2014. Long-read, whole-genome shotgun sequence data for five model organisms. *Scientific Data* **1**: 140045.
- Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution* **16**(2): 111-120.
- Kirschner J, Weber D, Neuschl C, Franke A, Bottger M, Zielke L, Powalsky E, Groth M, Shagin D, Petzold A et al. 2012. Mapping of quantitative trait loci controlling lifespan in the short-lived fish *Nothobranchius furzeri*--a new vertebrate model for age research. *Aging cell* **11**(2): 252-261.
- Koch P. 2010. Repetitive Elemente im Genom von *Nothobranchius furzeri* – Charakterisierung und Implikationen für die Genomsequenzierung (Diploma thesis). *unpublished*.
- Koch P, Platzer M, Downie BR. 2014. RepARK--de novo creation of repeat libraries from whole-genome NGS reads. *Nucleic acids research* **42**(9): e80.
- Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC bioinformatics* **7**: 474.
- Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED et al. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology* **30**(7): 693-700.
- Koressaar T, Remm M. 2012. Characterization of species-specific repeats in 613 prokaryotic species. *DNA research : an international journal for rapid publication of reports on genes and genomes* **19**(3): 219-230.
- Kramerov DA, Vassetzky NS. 2011. Origin and evolution of SINEs in eukaryotic genomes. *Heredity* **107**(6): 487-495.
- Krassovsky K, Henikoff S. 2014. Distinct chromatin features characterize different classes of repeat sequences in *Drosophila melanogaster*. *BMC genomics* **15**(1): 105.
- Krupovic M, Koonin EV. 2014. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nature reviews Microbiology*.

- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome research* **19**(9): 1639-1645.
- Kuhn RM, Haussler D, Kent WJ. 2013. The UCSC genome browser and associated tools. *Briefings in bioinformatics* **14**(2): 144-161.
- Lai Z, Jing J, Aston C, Clarke V, Apodaca J, Dimalanta ET, Carucci DJ, Gardner MJ, Mishra B, Anantharaman TS et al. 1999. A shotgun optical map of the entire Plasmodium falciparum genome. *Nature genetics* **23**(3): 309-313.
- Lander ES. 2011. Initial impact of the sequencing of the human genome. *Nature* **470**(7333): 187-197.
- Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**(4): 357-359.
- Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**(3): R25.
- Lee H, Gurtowski J, Yoo S, Marcus S, McCombie WR, Schatz M. 2014. Error correction and assembly complexity of single molecule sequencing reads.
- Lerat E. 2010. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* **104**(6): 520-533.
- Levels PJ, Gubbels REMB, Denucé JM. 1986. Oxygen consumption during embryonic development of the annual fish *Nothobranchius korthausae* with special reference to diapause. *Comparative Biochemistry and Physiology Part A: Physiology* **84**(4): 767-770.
- Lewin HA, Larkin DM, Pontius J, O'Brien SJ. 2009. Every genome sequence needs a good map. *Genome research* **19**(11): 1925-1928.
- Li R, Ye J, Li S, Wang J, Han Y, Ye C, Wang J, Yang H, Yu J, Wong G et al. 2005. ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Computational Biology* **preprint**(2005): e43.
- Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K et al. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20**(2): 265-272.
- Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**(13): 1658-1659.
- Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B et al. 2012. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Briefings in functional genomics* **11**(1): 25-37.
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(5950): 289-293.
- Lim KG, Kwok CK, Hsu LY, Wirawan A. 2013. Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. *Briefings in bioinformatics* **14**(1): 67-81.

- Long EO, Dawid IB. 1980. Repeated genes in eukaryotes. *Annual review of biochemistry* **49**: 727-764.
- Lopez-Flores I, Garrido-Ramos MA. 2012. The repetitive DNA content of eukaryotic genomes. *Genome dynamics* **7**: 1-28.
- Lucier JF, Perreault J, Noel JF, Boire G, Perreault JP. 2007. RTAnalyzer: a web application for finding new retrotransposons and detecting L1 retrotransposition signatures. *Nucleic acids research* **35**(Web Server issue): W269-274.
- Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**(1): 18.
- Macas J, Neumann P. 2007. Ogre elements--a distinct group of plant Ty3/gypsy-like retrotransposons. *Gene* **390**(1-2): 108-116.
- Mak AC, Lai YY, Lam ET, Kwok TP, Leung AK, Poon A, Mostovoy Y, Hastie AR, Stedman W, Anantharaman T et al. 2016. Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays. *Genetics* **202**(1): 351-362.
- Manning JE, Schmid CW, Davidson N. 1975. Interspersion of repetitive and nonrepetitive DNA sequences in the *Drosophila melanogaster* genome. *Cell* **4**(2): 141-155.
- Marcais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**(6): 764-770.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057): 376-380.
- Marie-Nelly H, Marbouty M, Cournac A, Flot JF, Liti G, Parodi DP, Syan S, Guillen N, Margeot A, Zimmer C et al. 2014. High-quality genome (re)assembly using chromosomal contact data. *Nature communications* **5**: 5695.
- Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**(2): 560-564.
- Merkel A, Gemmell N. 2008. Detecting short tandem repeats from genome data: opening the software black box. *Briefings in bioinformatics* **9**(5): 355-366.
- Metzker ML. 2005. Emerging technologies in DNA sequencing. *Genome research* **15**(12): 1767-1776.
- Metzker ML. 2010. Sequencing technologies - the next generation. *Nature reviews Genetics* **11**(1): 31-46.
- MHC Sequencing Consortium T. 1999. Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. *Nature* **401**(6756): 921-923.
- Mikheyev AS, Tin MM. 2014. A first look at the Oxford Nanopore MinION sequencer. *Molecular ecology resources* **14**(6): 1097-1102.
- Mirsky AE, Ris H. 1951. The desoxyribonucleic acid content of animal cells and its evolutionary significance. *The Journal of general physiology* **34**(4): 451-462.
- Morissette G, Flamand L. 2010. Herpesviruses and chromosomal integration. *Journal of virology* **84**(23): 12100-12109.

- Mullikin JC, Ning Z. 2003. The phusion assembler. *Genome research* **13**(1): 81-90.
- Myers EW. 2000. A Whole-Genome Assembly of *Drosophila*. *Science* **287**(5461): 2196-2204.
- Nadeau JH. 1989. Maps of linkage and syntenic homologies between mouse and man. *Trends in Genetics* **5**: 82-86.
- Nagarajan N, Pop M. 2013. Sequence assembly demystified. *Nature reviews Genetics* **14**(3): 157-167.
- Nagarajan N, Read TD, Pop M. 2008. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics* **24**(10): 1229-1235.
- Ng'oma E, Reichwald K, Dorn A, Wittig M, Balschun T, Franke A, Platzer M, Cellerino A. 2014. The age related markers lipofuscin and apoptosis show different genetic architecture by QTL mapping in short-lived *Nothobranchius* fish. *Aging-Us* **6**(6): 468-480.
- Niimura Y, Nei M. 2003. Evolution of olfactory receptor genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **100**(21): 12235-12240.
- Novak P, Neumann P, Pech J, Steinhaisl J, Macas J. 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**(6): 792-793.
- Ohno S. 1972. So much "junk" DNA in our genome. *Brookhaven symposia in biology* **23**: 366-370.
- Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* **284**(5757): 604-607.
- Ouyang S, Buell CR. 2004. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic acids research* **32**(Database issue): D360-363.
- Passarge E, Horsthemke B, Farber RA. 1999. Incorrect use of the term syntenic. *Nature genetics* **23**(4): 387-387.
- Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, Stutz AM, Stedman W, Anantharaman T, Hastie A et al. 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods* **12**(8): 780-786.
- Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. *Trends in genetics : TIG* **17**(1): 23-28.
- Petzold A, Reichwald K, Groth M, Taudien S, Hartmann N, Priebe S, Shagin D, Englert C, Platzer M. 2013. The transcript catalogue of the short-lived fish *Nothobranchius furzeri* provides insights into age-dependent changes of mRNA levels. *BMC genomics* **14**: 185.
- Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America* **98**(17): 9748-9753.
- Phillippy AM, Schatz MC, Pop M. 2008. Genome assembly forensics: finding the elusive mis-assembly. *Genome biology* **9**(3): R55.
- Platt RN, 2nd, Blanco-Berdugo L, Ray DA. 2016. Accurate transposable element annotation is vital when analyzing new genome assemblies. *Genome biology and evolution*.
- Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**: i351-358.

- Putnam NH, O'Connell B, Stites JC, Rice BJ, Fields A, Hartley PD, Sugnet CW, Haussler D, Rokhsar DS, Green RE. 2015. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *arXivorg*.
- Quesneville H, Nouaud D, Anxolabehere D. 2003. Detection of new transposable element families in *Drosophila melanogaster* and *Anopheles gambiae* genomes. *Journal of molecular evolution* **57 Suppl 1**: S50-59.
- Quick J, Quinlan AR, Loman NJ. 2014. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *GigaScience* **3**: 22.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6): 841-842.
- Reichwald K, Lauber C, Nanda I, Kirschner J, Hartmann N, Schories S, Gausmann U, Taudien S, Schilhabel MB, Szafranski K et al. 2009. High tandem repeat content in the genome of the short-lived annual fish *Nothobranchius furzeri*: a new vertebrate model for aging research. *Genome biology* **10**(2): R16.
- Reichwald K, Petzold A, Koch P, Downie BR, Hartmann N, Pietsch S, Baumgart M, Chalopin D, Felder M, Bens M et al. 2015. Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish. *Cell* **163**(6): 1527-1538.
- Reinert K, Langmead B, Weese D, Evers DJ. 2015. Alignment of Next-Generation Sequencing Reads. *Annual review of genomics and human genetics* **16**: 133-151.
- Ren Y, Zhao H, Kou Q, Jiang J, Guo S, Zhang H, Hou W, Zou X, Sun H, Gong G et al. 2012. A high resolution genetic map anchoring scaffolds of the sequenced watermelon genome. *PloS one* **7**(1): e29453.
- Ribeiro FJ, Przybylski D, Yin S, Sharpe T, Gnerre S, Abouelleil A, Berlin AM, Montmayeur A, Shea TP, Walker BJ et al. 2012. Finished bacterial genomes from shotgun sequence data. *Genome research* **22**(11): 2270-2277.
- Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry* **242**(1): 84-89.
- Ronaghi M, Uhlen M, Nyren P. 1998. A sequencing method based on real-time pyrophosphate. *Science* **281**(5375): 363, 365.
- Rondeau EB, Minkley DR, Leong JS, Messmer AM, Jantzen JR, von Schalburg KR, Lemon C, Bird NH, Koop BF. 2014. The genome and linkage map of the northern pike (*Esox lucius*): conserved synteny revealed between the salmonid sister group and the Neoteleostei. *PloS one* **9**(7): e102089.
- Rudd MK, Wray GA, Willard HF. 2006. The evolutionary dynamics of alpha-satellite. *Genome research* **16**(1): 88-96.
- Sackton TB, Kulathinal RJ, Bergman CM, Quinlan AR, Dopman EB, Carneiro M, Marth GT, Hartl DL, Clark AG. 2009. Population genomic inferences from sparse high-throughput sequencing of two populations of *Drosophila melanogaster*. *Genome biology and evolution* **1**: 449-465.
- Saha S, Bridges S, Magbanua ZV, Peterson DG. 2008a. Computational Approaches and Tools Used in Identification of Dispersed Repetitive DNA Sequences. *Tropical Plant Biology* **1**(1): 85-96.
- Saha S, Bridges S, Magbanua ZV, Peterson DG. 2008b. Empirical comparison of ab initio repeat finding programs. *Nucleic acids research* **36**(7): 2284-2294.

References

- Salmela L, Makinen V, Valimaki N, Ylinen J, Ukkonen E. 2011. Fast scaffolding with small independent mixed integer programs. *Bioinformatics* **27**(23): 3259-3265.
- Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M et al. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome research* **22**(3): 557-567.
- Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**(12): 5463-5467.
- Schartl M. 2014. Beyond the zebrafish: diverse fish species for modeling human disease. *Disease models & mechanisms* **7**(2): 181-192.
- Schartl M, Walter RB, Shen Y, Garcia T, Catchen J, Amores A, Braasch I, Chalopin D, Volff JN, Lesch KP et al. 2013. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. *Nature genetics* **45**(5): 567-572.
- Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. *Genome research* **20**(9): 1165-1173.
- Scheicher RH, Grigoriev A, Ahuja R. 2012. DNA sequencing with nanopores from an ab initio perspective. *Journal of Materials Science* **47**(21): 7439-7446.
- Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK. 1993. Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* **262**(5130): 110-114.
- Sharma SK, Bolser D, de Boer J, Sonderkaer M, Amoros W, Carboni MF, D'Ambrosio JM, de la Cruz G, Di Genova A, Douches DS et al. 2013. Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. *G3* **3**(11): 2031-2047.
- Shearer LA, Anderson LK, de Jong H, Smit S, Goicoechea JL, Roe BA, Hua A, Giovannoni JJ, Stack SM. 2014. Fluorescence in situ hybridization and optical mapping to correct scaffold arrangement in the tomato genome. *G3* **4**(8): 1395-1405.
- Shelton JM, Coleman MC, Herndon N, Lu N, Lam ET, Anantharaman T, Sheth P, Brown SJ. 2015. Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC genomics* **16**(1): 734.
- Simpson JT, Durbin R. 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome research* **22**(3): 549-556.
- Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome research* **19**(6): 1117-1123.
- Smit AFA, Hubley R. 2008-2010. RepeatModeler Open-1.0 <http://www.repeatmasker.org>.
- Smit AFA, Hubley R, Green P. 1996-2015. RepeatMasker Open-4.0 <http://www.repeatmasker.org>.
- Smith CD, Shu S, Mungall CJ, Karpen GH. 2007. The Release 5.1 annotation of *Drosophila melanogaster* heterochromatin. *Science* **316**(5831): 1586-1591.
- Staden R. 1979. A strategy of DNA sequencing employing computer programs. *Nucleic acids research* **6**(7): 2601-2610.

- Staton SE, Burke JM. 2015. Transposome: a toolkit for annotation of transposable element families from unassembled sequence reads. *Bioinformatics* **31**(11): 1827-1829.
- Tang H. 2007. Genome assembly, rearrangement, and repeats. *Chemical reviews* **107**(8): 3391-3406.
- Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, Schnable PS, Lyons E, Lu J. 2015. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome biology* **16**(1).
- Terzibasi E, Lefrancois C, Domenici P, Hartmann N, Graf M, Cellerino A. 2009. Effects of dietary restriction on mortality and age-related phenotypes in the short-lived fish *Nothobranchius furzeri*. *Aging cell* **8**(2): 88-99.
- Terzibasi E, Valenzano DR, Benedetti M, Roncaglia P, Cattaneo A, Domenici L, Cellerino A. 2008. Large differences in aging phenotype between strains of the short-lived annual fish *Nothobranchius furzeri*. *PloS one* **3**(12): e3866.
- Thomas CA, Jr. 1971. The genetic organization of chromosomes. *Annual review of genetics* **5**: 237-256.
- Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews Genetics* **13**(1): 36-46.
- Valdesalici S, Cellerino A. 2003. Extremely short lifespan in the annual fish *Nothobranchius furzeri*. *Proceedings Biological sciences / The Royal Society* **270 Suppl 2**: S189-191.
- Valenzano DR, Benayoun BA, Singh PP, Zhang E, Etter PD, Hu CK, Clement-Ziza M, Willemssen D, Cui R, Harel I et al. 2015. The African Turquoise Killifish Genome Provides Insights into Evolution and Genetic Architecture of Lifespan. *Cell* **163**(6): 1539-1554.
- Valenzano DR, Kirschner J, Kamber RA, Zhang E, Weber D, Cellerino A, Englert C, Platzer M, Reichwald K, Brunet A. 2009. Mapping loci associated with tail color and sex determination in the short-lived fish *Nothobranchius furzeri*. *Genetics* **183**(4): 1385-1395.
- Valenzano DR, Terzibasi E, Cattaneo A, Domenici L, Cellerino A. 2006a. Temperature affects longevity and age-related locomotor and cognitive decay in the short-lived fish *Nothobranchius furzeri*. *Aging cell* **5**(3): 275-278.
- Valenzano DR, Terzibasi E, Genade T, Cattaneo A, Domenici L, Cellerino A. 2006b. Resveratrol prolongs lifespan and retards the onset of age-related markers in a short-lived vertebrate. *Current biology : CB* **16**(3): 296-300.
- Valouev A, Zhang Y, Schwartz DC, Waterman MS. 2006. Refinement of optical map assemblies. *Bioinformatics* **22**(10): 1217-1224.
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al. 2001. The sequence of the human genome. *Science* **291**(5507): 1304-1351.
- Volff JN, Bouneau L, Ozouf-Costaz C, Fischer C. 2003. Diversity of retrotransposable elements in compact pufferfish genomes. *Trends in genetics : TIG* **19**(12): 674-678.
- Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, Lathrop M. 1992. A second-generation linkage map of the human genome. *Nature* **359**(6398): 794-801.
- Wetzel J, Kingsford C, Pop M. 2011. Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. *BMC bioinformatics* **12**: 95.

References

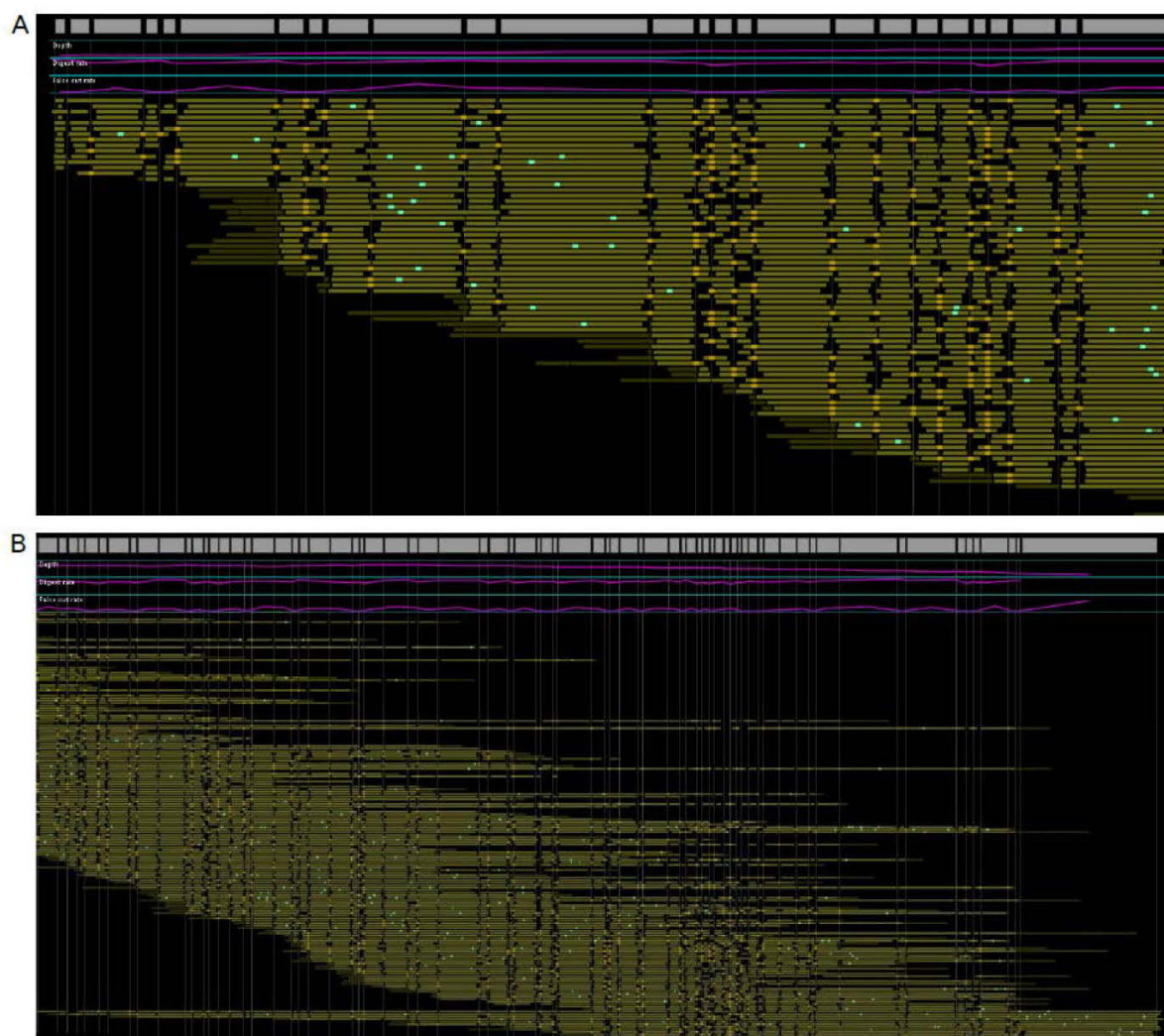
- Wicker T, Matthews DE, Keller B. 2002. TREP: a database for Triticeae repetitive elements. *Trends in Plant Science* **7**(12): 561-562.
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O et al. 2007. A unified classification system for eukaryotic transposable elements. *Nature reviews Genetics* **8**(12): 973-982.
- Wilson VL, Smith RA, Ma S, Cutler RG. 1987. Genomic 5-methyldeoxycytidine decreases with age. *The Journal of biological chemistry* **262**(21): 9948-9951.
- Xie T, Zheng JF, Liu S, Peng C, Zhou YM, Yang QY, Zhang HY. 2015. De novo plant genome assembly based on chromatin interactions: a case study of *Arabidopsis thaliana*. *Molecular plant* **8**(3): 489-492.
- Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nature reviews Genetics* **13**(5): 329-342.
- Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**(5): 821-829.
- Zhou S, Deng W, Anantharaman TS, Lim A, Dimalanta ET, Wang J, Wu T, Chunhong T, Creighton R, Kile A et al. 2002. A Whole-Genome Shotgun Optical Map of *Yersinia pestis* Strain KIM. *Applied and Environmental Microbiology* **68**(12): 6321-6331.
- Zhou S, Wei F, Nguyen J, Bechner M, Potamousis K, Goldstein S, Pape L, Mehan MR, Churas C, Pasternak S et al. 2009. A single molecule scaffold for the maize genome. *PLoS genetics* **5**(11): e1000711.

Supplemental Figures

```
Let  $\mathbb{E}$  be all edges in descending reverse sorted distance
For edge  $e_i(\text{SourceNode}_{e_i}, \text{DestNode}_{e_i})$  in  $\mathbb{E}$ 
    If node  $\text{SourceNode}_{e_i}$  has  $> 1$  edges  $\mathbb{E}_i$  in the same direction
        For  $e_j$  in  $\mathbb{E}_i$ 
            If  $\text{Size}(\text{DestNode}_{e_j}) < \text{Size}(\text{SourceNode}_{e_i})$ 
                Create edge  $e_x(\text{DestNode}_{e_j}, \text{DestNode}_{e_i})$  if it does not exist
                Delete  $e_i$ 
            Else if  $\text{Size}(\text{DestNode}_{e_j}) < \text{Size}(\text{Insert Library})$ 
                Delete  $e_i$ 
            Else
                Delete all edges in  $\mathbb{E}_i$ 
```

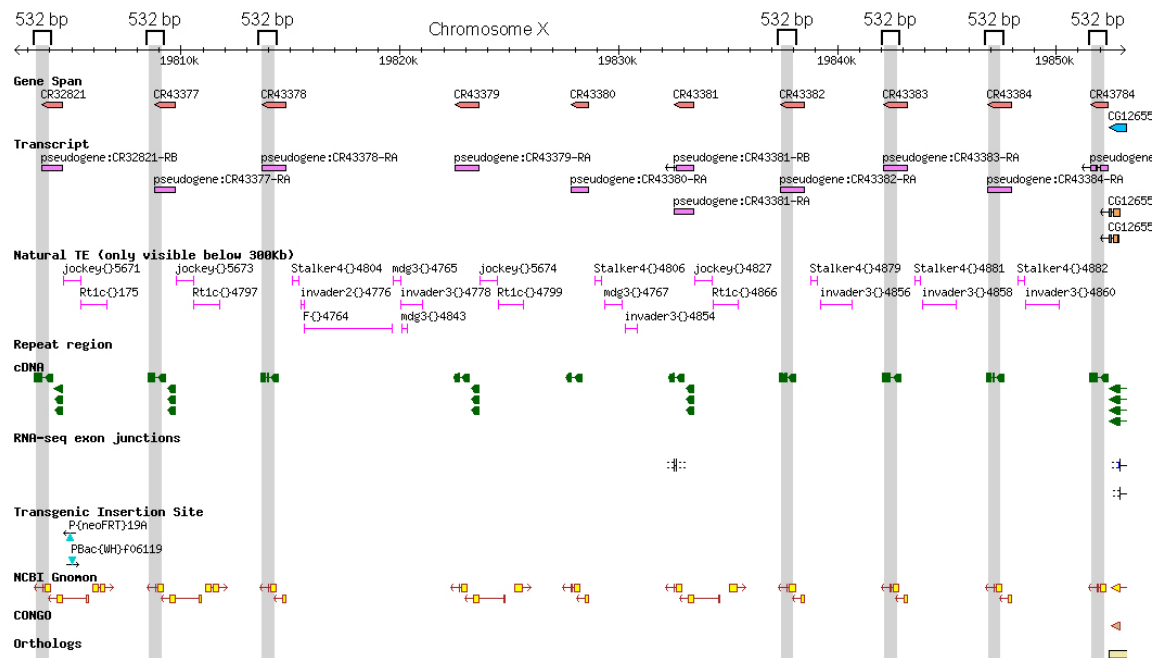
Supplemental Figure 1: Edge Pruning Algorithm of KILAPE.

During scaffolding, a graph of contig pairing (including gap size estimation and relative contig orientation) is built based on anchored paired-reads which link contigs. A graph is constructed where nodes are the contig IDs and the edges contain information about predicted distance and node direction. Edges are pruned heuristically from the graph in order of descending. This algorithm ensures that each contig has at most one edge in each direction, eliminating pairing discrepancies and providing a graph solution. [B. Downie, personal communication]



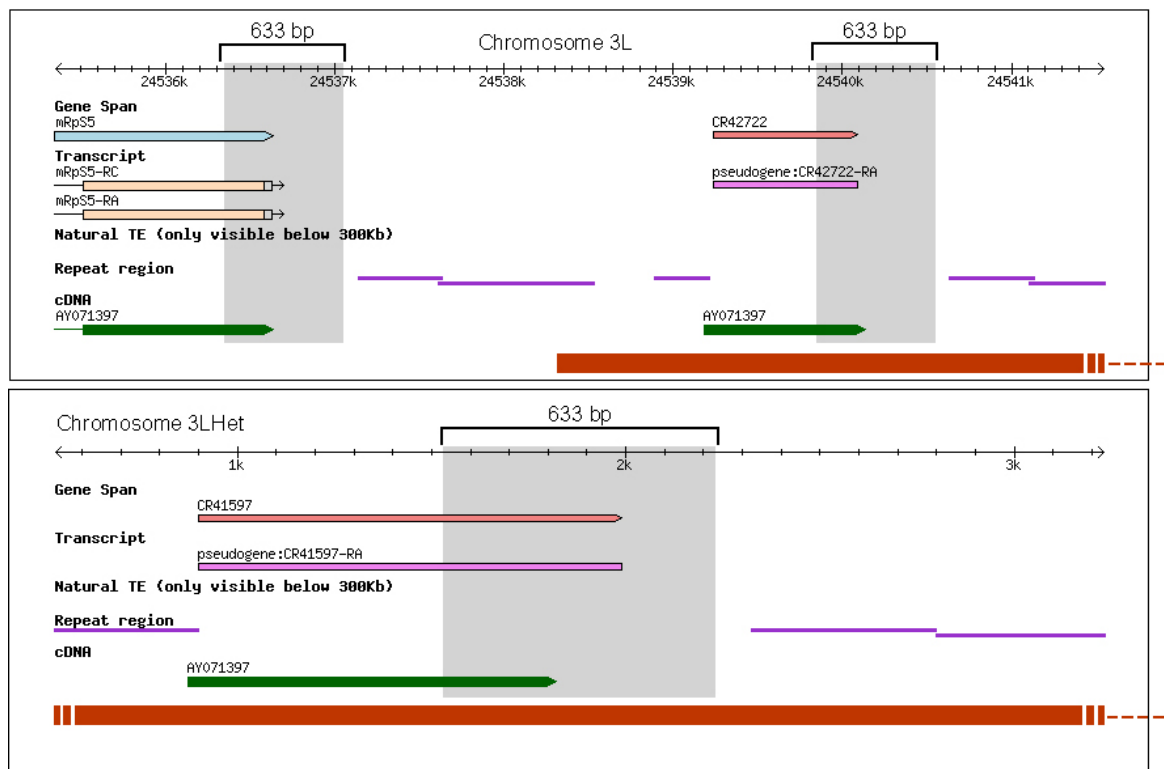
Supplemental Figure 2: SMRM Alignments at the Ends of superscaffold00011.

(A) left maptig end with a SMRM coverage of 13-fold. (B) right maptig end with a SMRM coverage of 12-fold.



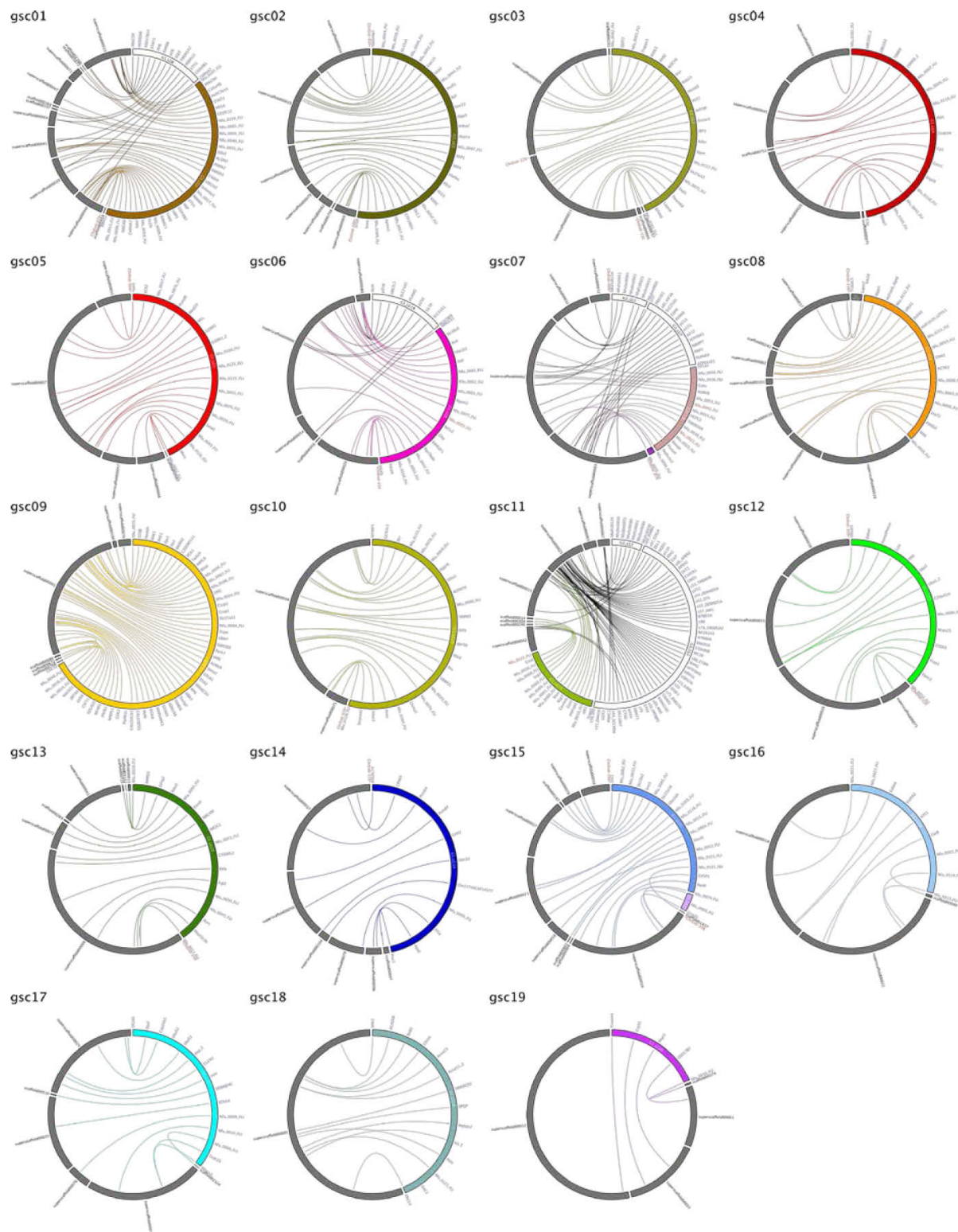
Supplemental Figure 3: Example of Short Length Putative SD Event in *D. melanogaster*.

A 532 bp consensus from the real RepARK CLC library shows a pattern of SD (gray shaded) on chromosome X (100% identity). The snapshot was taken from flybase.org; chromosomal location and consensus lengths were added. [from Figure S4 (Koch et al. 2014)]



Supplemental Figure 4: Example of SDs Involving Heterochromatin Identified in *D. melanogaster*.

A consensus from the simulated RepARK CLC library shows a pattern of SD (gray shaded) within chromosome 3L. This particular consensus also maps to the 3L heterochromatin. Red bar (4.4 kb) represents a pair of a SD detected by Eichler et al. The snapshot was taken from flybase.org; chromosomal locations, consensus lengths and SD annotation were added. [from Figure S5 (Koch et al. 2014)]



Supplemental Figure 5: Genetic Scaffold Construction.

Scaffolds and superscaffolds from assembly C are depicted in dark gray at the left side of each circle and represent their final order within each GSC. LGs with their respective markers are at the right side of the circle. LGs filled with different colors represent the linkage map G1 while the non-filled groups belong to G2 and G3. LGs are named according to their affiliation to the three maps as “G1_LG2” means LG2 of the map G1. Lines within the circles represent alignments of markers to the scaffolds and superscaffolds. Chromosome end annotations are labeled in red. For convenience/simplicity, markers of the genetic maps are depicted as evenly distributed and do not reflect their real distances.

Supplemental Tables

Supplemental Table 1: Statistics of Whole Genome *de novo* Assemblies of *D. melanogaster*.

	simulated reads		real reads	
Assembler	Velvet	wgs-assembler	Velvet	wgs-assembler
Number of sequences	66,720	1,348	47,680	4,045
Total length [Mb]	126.7	121.8	117.7	116.7
Min / max length	57 bp / 1.3 Mb	103 bp / 1.4 Mb	57 bp / 0.3 Mb	166 bp / 434 kb
Average [kb]	1.9	90	2.5	29
N50 [kb]	276	328	9	73
N90 [kb]	19	54	4	13

[modified from Table S1 (Koch et al. 2014)]

Supplemental Table 2: Summary of Microsatellite and SNP Markers of *N. furzeri* Genetic Map 1 (G1) Assigned to 19 LGs.

Linkage group	Number of markers	Number of markers that map to a scaffold or superscaffold
LG01	36	36
LG02	30	30
LG03	24	24
LG04	18	18
LG05	20	20
LG06	20	19
LG07	17	14
LG08	18	18
LG09	57	57
LG10	24	24
LG11	17	16
LG12	14	14
LG13	17	17
LG14	11	11
LG15	19	19
LG16	9	9
LG17	15	15
LG18	14	14
LG19	2	2
LG20	5	5
LG21	3	3
LG22	2	2
Total	392	387

^a Ratio to the total number of consensi of the library. ^b Ratio to the total length of the library, N/A: not applicable

Supplemental Table 3: Repeat Consensi of *D. melanogaster* Classified with TEclass.

	Library	Consensi analysed ^a	DNA transposons	Retrotransposons	Not classified
Sanger	DmRepBase	249	20.5%	77.9%	1.6%
	ReASLib	391	25.8%	67.3%	6.9%
Simulated reads	RepeatScout	35,043 ^b	53.7%	38.0%	8.3%
	wgs-assembler	14,147	43.7%	47.7%	8.6%
	RepARK CLC	1,239	38.6%	53.7%	7.7%
	RepARK Velvet	18,203	23.3%	69.4%	7.3%
Real reads	RepeatScout	11,439 ^c	46.4%	46.3%	7.3%
	wgs-assembler	4,284	32.5%	60.8%	6.7%
	RepARK CLC	414	47.3%	44.2%	8.5%
	RepARK Velvet	14,296	35.9%	56.7%	7.4%

^a Number of consensi ≥ 50 bp in the respective libraries, which refers to all consensi of DmRepBase, ReASLib, wgs-assembler, RepeatScout, simulated and real RepARK Velvet libraries. ^b 52% of consensi in simulated RepARK CLC and 71% (3 Mb) of the overall length. ^c 53% of consensi in real RepARK CLC and 80% (1.3 Mb) of the overall length. [modified from Table S4 (Koch et al. 2014)]

Acknowledgements

I am grateful to PD Dr. Matthias Platzter, for giving me the opportunity to work on this interesting project and for his continuous support and advice. I also would like to thank Dr. Kathrin Reichwald, for her excellent mentoring, guidance and support. I thank Dr. Andreas Petzold for his invaluable support and encouragement in the daily Bioinformatics work. Additionally, I thank all the three of them for their help during the formation process and the critical reading of my dissertation.

I would like to thank all former and current colleagues of the Genome Analysis Group for the very friendly and constructive working environment, especially: Maja Kinga Dziegielewska, Silke Förste, Ivonne Görlich, Jeanette Haink, Ivonne Heinze, Cornelia Luge, Patricia Möckel, Stefanie Stepanow, Beate Szafranski, Martin Bens, Marius Felder, Marco Groth, Klaus Huse, Niels Jahn, Arne Sahm, Bernd Senf, Karol Szafranski and Stefan Taudien.

Moreover, I would like to thank my thesis committee members: Matthias Platzter, Jürgen Sühnel and Sabastian Böcker from the Friedrich-Schiller-University in Jena.

I thank Enoch Ng'oma for providing an additional linkage map that supported the genome assembly.

I thank Bryan Downie for his supporting and mentoring during the development of RepARK.

Furthermore, I would like to thank Jean-Nicolas Volff who enabled my stay at his lab, Domitille Chalopin Fillot for her valuable help and advices on the repeat classification and Christoph Englert who initiated this collaboration.

Lastly, I want to thank my entire family as well as the friends and colleagues who have not been mentioned here personally for their continuous encouragement to make this educational process a success. Surely I could not have made it without your supports.

Thank you, my love.

Selbstständigkeitserklärung

Ich erkläre, dass mir die geltende Promotionsordnung der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität in Jena bekannt ist. Ich versichere, dass ich die vorliegende Dissertation selbst angefertigt, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle verwendeten Hilfsmittel, persönlichen Mitteilungen und Quellen in meiner Arbeit angegeben habe.

Insbesondere waren folgende Personen in genannter Art und Weise direkt an der Entstehung der vorliegenden Arbeit beteiligt:

- Dr. Kathrin Reichwald, Dr. Marco Groth und Dr. Stefan Taudien (FLI Jena) bereiteten DNA-Proben vor und führten die Sanger, 454, Illumina und PacBio Sequenzierungen durch.
- Dr. Kathrin Reichwald (FLI Jena) stellte die Zellkultur für die Analyse durch OpGen bereit.
- Dr. Bryan Downie (FLI Jena) entwickelte KILAPE und hat damit die Assemblierung B durchgeführt.
- Dr. Andreas Petzold (FLI Jena) führte die Genannotation und Syntänieanalyse durch, half bei den Alignments der Kopplungskartenmarker und etablierte den *N. furzeri* Genombrowser.
- Dr. Domitille Chalopin Fillot (ENS Lyon) führte die Repeatanalyse mit RepeatModeler durch.

Ich bestätigte, dass ich die Hilfe eines Promotionsberaters nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die in Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Ich versichere, dass ich die Dissertation weder in gleicher noch in ähnlicher Form zuvor als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe.

Jena, 05.02.2016

.....

Philipp Koch